Studies in Educational Evaluation 43 (2014) 55–69

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/stueduc

Fairness as a multifaceted quality in classroom assessment

Robin D. Tierney [*]

University of British Columbia, MERM, Faculty of Education, Vancouver, British Columbia, Canada

ARTICLE INFO

Article history:
Received 15 May 2013
Received in revised form 8 November 2013
Accepted 17 December 2013
Available online 16 January 2014

Keywords:
Fairness
Quality assurance
Classroom assessment
Assessment for learning
Phronesis
Qualitative research

ABSTRACT

Fairness is an essential and enduring ideal in education, but it has not been clearly defined for the dynamics of classroom assessment. This paper aims to contribute to the reconceptualization of fairness as multifaceted quality in classroom assessment where the primary purpose is to support student learning. This multi-case study elicited the phronesis (practical wisdom) of six purposefully selected teachers in Ontario, Canada. They responded to fairness issues in written vignettes, and then discussed their concerns and gave recommendations for fair assessment during interviews. The participants emphasized different aspects of fairness with the most prominent involving students' opportunities to learn and demonstrate learning, transparency, the classroom environment, critical reflection, and the tension between equal and equitable treatment in assessment.

Quality assurance is a proactive process for determining and monitoring, either internally or externally, that a program or practice fulfills its purpose and meets stakeholder requirements (Doherty, 2008; Martin & Stella, 2007). Qualities that are desirable for educational assessment are usually identified in measurement theory as validity, reliability, and fairness (e.g., American Education Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 1999). These qualities can be considered for different levels of educational assessment, from individual learner diagnostics to system-wide accountability testing. In this paper, I focus specifically on reconceptualizing fairness as a multifaceted quality in classroom assessment (CA) that aims to support learning.

CA is an ongoing process that involves teachers and students in gathering information (assessing) and making judgments (evaluating) about student learning. CA results have traditionally been used to determine and report on achievement in order to place or certify students. This is referred to as summative assessment or assessment of learning (AofL). The use formative assessment or assessment *for* learning (AfL) has been increasingly endorsed in the educational assessment community (e.g., Assessment Reform Group [ARG], 1999; Earl & Katz, 2006; Stiggins & Chappuis, 2005). AfL involves sharing clear expectations and criteria, using varied methods to elicit learning, giving task-specific feedback to students, involving them in assessment processes, and using the results to inform teaching (Tierney & Charland, 2007). Wiliam (2011) emphasizes that for assessment to effectively support learning, it must provide specific information, not only to direct further teaching, but also to encourage student engagement in productive learning. The need for AfL is now broadly recognized (e.g., Gordon Commission, 2013; Ontario Ministry of Education, 2010), but little research has been done on fairness issues in its practice.

Going back and moving forward

During most of the 20th century, the educational measurement community focused on the development of standardized tests. To a large extent, this was because of the widely held belief that objectivity could be attained through the application of scientific technique (Clarke, Madaus, Horn, & Ramos, 2000), and inferences from test results were thought to have a higher degree of validity, reliability and fairness than CA results. As the century turned, shifting social ideals, evolving ideas about the nature of knowledge, developments in understanding human learning, and rapid technological advancements changed the educational landscape. The negative impact of high-stakes testing on teaching and learning was increasingly recognized (e.g., Abrams, Pedulla, & Madaus, 2003; Frederiksen & Collins, 1989; Gipps & Murphy, 1994; Heubert & Hauser, 1999; Popham, 2003), which not only heightened concerns about quality in testing, but also generated interest in developing and using large-scale performance assessments that would support student learning and measure achievement. The challenges of ensuring high technical quality in performance assessments quickly became apparent, and expert

[*] Tel.: +1 604 401 6217.
E-mail addresses: robin.tierney@ubc.ca, tierney@research-for-learning.com

0191-491X/$ – see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.stueduc.2013.12.003

voices repeatedly cautioned that an assessment should not be considered fairer simply because it aims to support learning (Elwood, 2006a; Gipps, 2005; Linn, Baker, & Dunbar, 1991).

Interest in the pedagogical potential of CA increased through this period, particularly with the emergence of research on the benefits for student learning (e.g., Black & Wiliam, 1998). In comparison to standardized testing, CA was commonly considered to be low-stakes. However, a growing body of research has shown that CA does affect student motivation and self-regulation in learning (Brookhart, 2013), and there can be long-term social, emotional, and academic consequences for students (Brookhart, 1994; Dalbert, Schneidewind, & Saalbach, 2007; Morgan & Watson, 2002). Research on the quality of 'newer' assessment methods shows mixed results (Dierick & Dochy, 2001), and some are potentially less fair than traditional tests because of their personal nature (e.g., reflective response journals) (Gynnild, 2011; Schendel and O'Neill, 1999). Recognizing the high-stakes of CA for learners brings us full circle to concerns about quality similar to those expressed by early 20th century edumetricians before the heyday of standardized testing (Finklestein, 1913; Monroe, 1923; Rinsland, 1937).

Efforts to assess the quality of CA from a measurement perspective have generally resulted in teachers' technical ability or assessment literacy being "found wanting" (Brookhart, 2004, p. 447). Rather than assuming that quality problems were caused entirely by poor practice, many assessment specialists at the end of the 20th century began to question the relevance of measurement theory for the dynamics of CA (Brookhart, 1993; Delandshere, 2001; Gipps & Murphy, 1994; Stiggins, Frisbie, & Griswold, 1989; Whittington, 1999; Wiggins, 1993; Wiliam, 1994). This fueled the development of two documents containing principles or standards for CA in North America (Joint Advisory Committee [JAC], 1993; Joint Committee on Standards for Educational Evaluation [JCSEE], 2003). The idea that CA should be fair is inherent to these documents, but neither explicitly defines the concept. This fuzziness around fairness remains to date despite the sustained thrust to re-conceptualize quality for CA. Considerable discussion pertains to validity and reliability (e.g., Black & Wiliam, 2006; Bonner, 2013; Moss, 2003; Parkes, 2007, 2013; Smith, 2003; Stobart, 2006), but much less focuses on fairness, and little guidance for fair AfL is given. A better understanding of fairness in CA is needed, especially for AfL with diverse students. I aim to contribute to that understanding with this multi-case study on teachers' phronesis about fairness. In the balance of this paper, I explain the rationale for my research approach, and I identify existing interpretations of fairness. I then describe the methodology and results, and link them back to existing literature in the discussion.

Rationale for turning to teachers for phronesis

For quality assurance in education, Doherty (2008) suggests that when "you want to improve something, ask the people who do it" (p. 82). Baartman, Bastiaens, Kirschner, and van der Vleuten (2006, 2007) took this approach when they consulted 12 international measurement experts to re-frame quality criteria for competency-based assessment programs, and subsequently surveyed Dutch vocational teachers regarding their framework. My research has a similar goal, with two main differences. First, I am concerned more specifically with AfL, which involves both planned events and spontaneous interactions (Cowie & Bell, 1999). Thus my interest in fairness extends beyond tests to the nebulous space between teachers, students and students' learning where inferences and decisions are made, often quickly and tacitly. Second, rather than seeking opinions on a framework developed by experts, I turned to teachers to shed light on a concept in their practices in order to improve the relevancy of CA theory.

The philosophical perspective underpinning this work is a form of critical pragmatism. Warms and Schroeder (1999/2009) describe pragmatism as a "way of doing philosophy that weaves together theory and action, each continuously modifying the other and maintain their mutual relevance" (p. 271). From this perspective, varied forms of knowledge are valuable for educational research (Biesta & Burbules, 2003; Maxcy, 2003), including ethical and practical knowledge. These are undervalued relative to theoretical and empirical knowledge (Fenstermacher, 1994; Kessels & Korthagen, 1996), and greater consideration should be given to the phronesis that guides practice (Bernstein, 1985; Biesta, 2007; Dunne, 1993). Phronesis is translated from classical Greek as practical wisdom, reasoning, or judgment (Dottori, 2009; Fenstermacher, 1994; Flyvbjerg, 2004). Aristotle defined it as a "reasoned and true state of capacity to act with regard to human goods" (Aristotle, trans. 1925/1975, p. 143). In contemporary terms, phronesis draws on a mental network that includes technical knowledge, theoretical knowledge, moral beliefs and professional ethics, personal characteristics, experience, and understanding of particulars (Connolly, 2005; Dunne, 1993). Essentially, it provides the ability "to judge correctly, all things considered, the best action to perform in any given situation" (Warne, 2006, p. 15). Phronesis shares characteristics with other conceptualizations of teachers' knowledge (e.g., Clandinin & Connelly, 1996; Shulman, 2004a), but it also includes a moral dimension that is key for research into the ethics of practice.

Interpretations of fairness in assessment theory

In the literature that aims to reconceptualize quality for CA, what or where we are moving from is not always clear, possibly because describing this is like painting a cloudy sky on a windy day. The meaning of fairness continues to evolve, along with other key qualities, in measurement theory. In early editions of testing standards (APA, AERA, & NCME, 1966, 1974), the terms *biased* and *unfair* were used interchangeably, but now fairness is recognized as a broader social concern that goes beyond technical issues of bias (Camilli, 2006; Moss, Pullin, Gee, & Haertel, 2005; Stobart, 2005). Four interpretations of fairness are discussed in the last edition (AERA, APA & NCME, 1999). The first two, being the absence of statistical bias and the equitable treatment of test takers in the testing process, have generally been accepted in the measurement community. The third interpretation, where fairness is associated with equality of test outcomes, is rejected based on the long-established point that group score differences do not necessarily indicate test bias (e.g., Messick & Anderson, 1970). The fourth interpretation, were fairness is associated with opportunity to learn is considered problematic, but it is acknowledged that prior access to test material is necessary for decisions based on test results to be fair. Discussion about fairness in testing is ongoing in the measurement community (e.g., responses to Xi, 2010). Nonetheless, there is general agreement on two points. First, fairness cannot be determined dichotomously because it is a matter of degree (Cole & Zieky, 2001; Lane & Silver, 1999). And second, it is an important quality that is distinct from, but related to validity (Camilli, 2006; Messick, 2000; Stobart, 2005).

While the meaning of fairness continues to evolve in testing, it is not defined in the principles and standards for CA (JAC, 1993; JCSEE, 2003), and research focusing specifically on fairness in CA is limited (Tierney, 2013). In preparation for this multi-case study, I culled existing interpretations from a range of texts (i.e., peer reviewed articles, CA textbooks, doctoral dissertations, and joint committee documents). Some interpretations conflict while others

Table 1
Operationalization of fairness for classroom assessment.

Steps[a]	Texts				
	Russell and Airasian (2012)	Camilli (2006)	McMillan (2011)	Suskie (2002)	Volante (2006)
1	Informing students about teacher expectations and assessments before beginning teaching and assessment	Clear and reasonable assessment criteria	Student knowledge of learning targets and assessments	Have clearly stated learning outcomes	Examine how your classroom complements or conflicts with the school experiences and assessment techniques which are familiar to recent immigrants and indigenous people
2	Teaching pupils what they are to be tested on before summative assessment	Equity in assessment and instruction [focused on learning]	Opportunity to learn	Match your assessment to what you teach and vice versa	Develop test questions and other assessment items that reflect the multicultural and multiethnic composition of your school, district, region, province, country
3	Not making snap judgments and identifying pupils with emotional labels (e.g., disinterested, at-risk, slow learner) before you have spent time with them	Opportunity to learn	Prerequisite knowledge and skills	Use many different measures and many different kinds of measures	Utilize gender-neutral terms within tests, quizzes, and other forms of assessment
4	Avoiding stereotyping pupils (e.g., "He's just a dumb jock," "Kids from that part of town are troublemakers," and "Pupils who dress that way have no interest in school")	Sensitivity and construction of assessments	Avoiding student stereotyping	Help students learn how to do the assessment task	Provide modifications to tests and other assessment measures for students with exceptionalities and those with limited proficiency in English
5	Avoiding terms and examples that may be offensive to students of different gender, race, religion, culture, or nationality	Multiple measures	Avoiding bias in assessment tasks or procedures	Engage and encourage your students.	Take steps to reduce the effects of test-wiseness on academic achievement – design carefully constructed test items, utilize multiple-formats, provide reviews prior to testing
6	Avoiding bias toward pupils with limited English or with different cultural experiences when providing instruction and constructing assessments	Modeling fairness	Accommodating special needs and English Language Learners	Interpret assessment results appropriately	Involve students in the development of evaluation criteria–develop rubrics with the assistance of students
7				Evaluate the outcomes of your assessments	Adopt a range of formative and summative assessment strategies that encompass different ways of demonstrating task mastery – speaking, writing, performing
8					Balance the weight given to different types of traditional and authentic performance-based assessment data when arriving at final course grades
9					Most importantly, reflect on preconceived notions that may affect the mark/grades you assign to particular groups of student

[a] Authors use the terms aspects, components, strategies, steps and practices.

Table 2
Case Essentials.

Pseudonym	Kevin	Lorena	York	Tanar	Lucy	Amada
Years of teaching	13	14	18	20	40	40
Current assignment	• G9 and 10 English, Applied • G9 and 10 English, Academic G12 English, University Prep	• G7 and 8 Special Education (English and Math) • Teacher-Librarian Literacy Lead Teacher	• G11 &12 English, College Prep G11 & 12 English, University Prep • G12 Geography	• G12 English, Writer's Craft (online across school board) Teacher-Librarian	• G10 &11 English, Cong Gifted G12 English, Writer's Craft	• Supply teaching, secondary level • Private tutoring, elementary level Retired teacher and school leader
Working environment	• Small secondary school (>400) • Military base, rural region Students: 91% speak English at home; OSSLT pass rate below provincial average	• Small intermediate school (>400) • Center of large urban area Students: 60% speak English at home, culturally diverse, wide range of abilities	• Very large learning center (<2000) • Inner-suburban ring of urban area • Students: Young adults aiming for skills and credits for secondary school diploma	• Medium-sized secondary school (>1000) with arts, AP and gifted programs • Suburban on urban boundary Students: 59% speak English at home; OSSLT pass rate above prov. avg.	• Large secondary school (<1000) with alternative, gifted, and IB programs • Suburban on urban outskirts Students: 34% speak English at home; OSSLT pass rate above prov. avg.	• Varied, include home and learning center with some credit courses • All in densely populated and culturally diverse area amalgamated through urban expansion
Practices identified as most important for fairness	• Providing clear learning expectations and assessment criteria • Designing assessment tasks for relevant opportunities to demonstrate learning • Using multiple and varied assessment to understand learning Maintaining consistency in treatment of students	• Differentiating appropriately for individual learners • Gathering info about learners from multiple sources • Reflecting with colleagues about assessment • Increasing the consistency of assessment across classrooms	• Involving learners in assessment (criteria development) • Using multiple assessments for informed decisions • Being aware of biases and balancing knowledge of learners with professional judgment	• Reflecting on use of knowledge about learners and learning to assess judiciously • According students the same opportunities while allowing for special circumstances Explicit teaching and modeling for a constructive orientation	• Identifying the assessment purpose to clearly communicate learning expectations • Maintaining a consistent standard while differentiating for individual learning needs Providing engaging opportunities to learn while demonstrating learning	• Praising equally while giving honest and accurate feedback • Designing appropriate assessment tasks and criteria (T knowledge emphasized) Orienting assessment constructively to support high expectations for ongoing learning

overlap, and they vary considerably in their scope and focus on fairness.[1] I describe how I used these interpretations further in the methods section. Here I look specifically at five that offer operational definitions because quality assurance requires both conceptualization and operationalization (Martin & Stella, 2007). Camilli (2006) discusses CA at the end of a chapter on fairness in testing. In addition to recommending six practices, he notes the importance of assessment purpose and he emphasizes the necessity of differentiation (p. 248). Russell and Airasian (2012) relate fairness to the ethical standards that should generally guide teacher–student interactions (e.g., respect, honesty) in a textbook for K-12 teacher education. They also identify six aspects of fairness, which they discuss in terms of teachers' responsibility and consequences for students (p. 22). McMillan (2011) includes a chapter on quality in his textbook for K-12 teacher education, and he identifies and explains six key components for fairness (pp. 80–86). Suskie (2002) recommends seven steps for fair assessment by professors in higher education (pp. 3–4). Volante (2006) suggests nine points for consideration in a faculty-published magazine for educators. Three similarities are evident across these definitions. First, they all associate fairness with opportunities for students to learn and/or demonstrate learning, which they interpret more broadly than in the testing standards (AERA, APA, & NCME, 1999). Second, the need for transparency, particularly in communicating about criteria, is identified in four definitions. Transparency is also discussed in the standards for CA (JAC, 1993; JCSEE, 2003) and for testing, but not specifically as a fairness issue. Finally, most of these definitions go beyond the idea of simply avoiding bias to recommend reflection about stereotypes, values, and biases in assessment practices. These similarities are reflective of tends in thinking about fairness across varied types of literature relating to CA. I return to these definitions in the next sections (Table 1).

Research design and methods

I chose a multi-case study design for this work because its purpose was to illuminate a complex concept in practice. The cases are instrumental in that their value is not purely intrinsic; they are used to "provide insight into an issue" (Stake, 2005, p. 445). Case study research is well-suited for questions that emerge from "everyday practice" (p. 29) because of its "particularistic, descriptive, and heuristic" nature, which allows it to "bring about the discovery of new meaning, extend the reader's experience, or confirm what is known" (Merriam, 1998, p. 30). Two questions central to this paper are: what phronesis about fair CA can teachers offer, and how might it contribute to the reconceptualization of fairness in CA? Each case represents an English-language teacher working with secondary students in Ontario, Canada. Salient features of the educational context are described below to inform transferability to other contexts. Details about the participants, the data collections tools, the strategies used for data analysis and quality assurance are then provided.

Educational context

Education is publicly funded in Ontario from Junior Kindergarten to Grade 12. The Ontario Ministry of Education (OME) governs 72 district school boards of considerable variety. Small rural northern boards have fewer than 1000 students, while the largest urban board has more than 250,000 students located in the densely populated south-eastern corridor, one of the most culturally diverse areas in the world (OECD, 2011). School boards also differ in religious orientation (i.e., non-denominational or Catholic) and instructional language (i.e., English and/or French), and many specialized schools and programs are offered (e.g., arts-based). Following Kindergarten and elementary education (Grades 1–6, ages 5–12), students begin their secondary education, which consists of two intermediate years (Grades 7–8, ages 12–14) and four secondary years (Grades 9–12, ages 14–18). Students are placed in academic or applied courses in Grades 9 and 10, which streams them into workplace, college or university preparation courses for their final two years.

Standardized report cards and province-wide testing were introduced in the 1990s as part of the reform to equalize students' educational experiences through centralization. Students are required to pass the Ontario Secondary School Literacy Test (OSSLT) for graduation. While maintaining these accountability features, the OME's revised assessment policy document emphasizes the need for differentiated instruction and fair assessment, and it explicitly states that the primary purpose of CA is to improve learning for "every student" (2010). The first version of this document was in circulation during this study (2008), but teachers' assessment literacy still depended to a large extent on their own initiatives or local leadership. At the time, some teachers were sufficiently critical of the OME's new assessment policies that public controversy ensued, particularly about grading in secondary classrooms (Laucius, 2009). As such, this study was set in an educational context where assessment reform was underway, but central polices were not universally accepted and teachers' practices were in flux.

Participants

Careful selection of participants was critical given my intended use of teachers' phronesis. It is one of the "embarrassments of practice" (p. 264), as Shulman (2004b) points out, that not all practitioners are wise and "wise practitioners vary" (p. 265). For this reason and to amplify the utility of a small number of cases, I purposefully selected participants (Flyvbjerg, 2006; Patton, 2002; Stake, 2005). I determined the characteristics of teachers who would be most likely to share phronesis and provide rich information by conducting secondary analyses of pan-Canadian data from a teachers' questionnaire administered as part of a large-scale student assessment program (Council of Ministers of Education Canada [CMEC], 2002). Specifically, I examined the relationship between years of teaching experience, type of degree, and level of specialization in teaching English Language Arts, and I concluded that I needed teachers who had at least 10 years of teaching experience, held a relevant undergraduate degree, and identified themselves as specialists in teaching English Language Arts. Teachers who met these criteria and who volunteered on a teachers' network to help others with questions about CA were contacted by email. The six participants selected had between 13 and 40 years of teaching experience at the time (see Table 2). In addition to basic certification by the Ontario College of Teachers (OCT), they all had advanced qualifications for teaching English Language Arts in secondary grades. Although they were all located in the same educational system, they worked with a wide range of learners in different school environments.

Research methods

Two written questionnaires and follow-up interviews were used to gather information from the participants. The first questionnaire focused on the participants' experience and education (multiple-choice, dichotomous items), and it also included short items on the ethics of CA (based on Green, Johnson, Kim, & Pope, 2007). The second questionnaire contained eight written vignettes ranging from 92 to 107 words each. Vignettes are especially well-suited for eliciting responses from participants in research on ethical dilemmas or sensitive topics (Barter & Renold, 2000; Hughes &

[1] For a complete list of literature reviewed, see Tierney (2008).

Huby, 2004; Wilks, 2004). I created a blueprint for 15 vignettes with each describing an AfL practice (Cowie & Bell, 1999; Tierney & Charland, 2007) that was situated in a commonly used (CMEC, 2002) and officially mandated writing activity (OME, 2007). The resulting vignettes were reviewed by two established scholars in CA, two secondary teachers with relevant experience in Ontario, and two doctoral colleagues. Timed responses suggested that the questionnaire should be shortened to eight vignettes, and I revised the most relevant for the study's purpose based on reviewers' comments. The final vignettes included fairness issues relating to opportunity to learn (3), transparency (3), avoiding bias (3), and either equal or equitable treatment (3). Each vignette was followed by open-ended questions that asked participants to recommend the *fairest* decision for the dilemma. Follow-up interviews ranged from 50 to 75 min, during which I asked for clarifications and probed written responses. A benefit of vignettes is that they can be used not only to elicit responses related directly to their content, but also to prompt the sharing of personal experiences (Barter & Renold, 2000). All six participants described fairness issues they had experienced or dealt with in their own practices. I aimed for a collegial tone throughout to encourage frank discussion about fair/unfair assessment, and written feedback from participants suggests that this was achieved.

Data analyses were both deductive and inductive. Deductive analysis involves matching the participants' responses to categories that are established beforehand (Patton, 2002; Yin, 2006). I began by coding the participants' written responses and the interview transcripts for variations of the facets of fairness represented in the vignettes:

- Avoiding bias (role of values, stereotypes, power dynamics)
- Equal treatment of students (consistent content, criteria, processes)
- Equitable treatment of students (varied content, processes; differentiated for individuals)
- Opportunity to learn (alignment with teaching; adequate time; pre-requisite knowledge)
- Transparency (explicit and clear expectations, criteria, instructions, processes, results)

To extend beyond confirmation and gain insight into the participants' own understanding of fairness, I used a constant comparative-type method where categories emerge and patterns become evident as the analysis proceeds (Glaser & Strauss, 1967; also Merriam, 1998). I also looked at emerging themes in relation to additional issues or characteristics that are associated in varying degrees with fairness in existing literature:

- Appropriate (developmentally)
- Balanced (both strengths and weaknesses)
- Consequences of the assessment considered
- Constructive (focusing on learning)
- Focused content (feasible number of criteria)
- Guided by written policy and procedures
- Multiple evaluators (team teachers, peers, external)
- Opportunity to demonstrate learning
- Privacy of individual results protected
- Recognition of social values, cultural expectations in content and criteria
- Relevant (separation of academic and behavioral)
- Results carefully summarized (technical, meaningful)
- Results interpreted in sensitive manner (contextual, ipsative, regard for limitations)
- Respect for and sensitivity to students' needs (emotional, cultural, intellectual)
- Student participation in assessment

Through this process I moved between cases, working iteratively between newly and previously coded sections. This allowed me to eliminate, amalgamate and expand on pre-existing categories, and to continue clarifying themes.

I used multiple strategies and sources to support the quality of this work, beginning with an in-depth review of the context, and careful development of data collection tools. To minimize circularity and avoid constraining or leading participants' responses with the vignettes, I used open-ended response options, sensitive but probing follow-up questions, and complementary methods (Barter & Renold, 2000; Wilks, 2004). I also followed recommended practices for qualitative research, including verbatim transcription (Maxwell, 2005; Seidman, 1998) and triangulation of data (Erzberger & Kelle, 2003; Merriam, 1998; Stake, 2005; Yin, 2006), and I engaged in critical discussions about this project with my advisor and colleagues.

Phronesis about fairness in classroom assessment

The participants had a great deal to say about fairness, not only in response to the vignettes, but also regarding their own assessment practices. Because they all discussed multiple aspects of fairness, they were asked to identify which were *most* important and provide recommendations for other teachers and researchers to consider.

Kevin: the same hoop

Kevin had been a teacher for 13 years. He held a Bachelors in Arts (Drama) and Education, had additional qualifications in Dramatic Arts and English, and he had engaged in professional development in educational assessment. Kevin taught students in Grades 9, 10 and 12 English at a small secondary school (<400 students) on a military base in a rural area. While almost all of the students spoke English as their first language, many were in applied courses, and the overall achievement rate on a provincial literacy test was below average. Kevin aimed to be a good role model because he felt that many of his students did not "have anybody to look after them and teach them right from wrong" while their parents were in active service.

Kevin emphasized the need for routine and discipline in his classroom, but his comments also suggested that he was flexible and caring in his interactions with individual students. He emphasized four aspects of fairness: (1) using clear learning expectations and assessment criteria, (2) designing assessment tasks to provide relevant opportunities for students to demonstrate learning, (3) using varied assessments tasks to inform his understanding of student learning, and (4) equal treatment in assessment. Kevin aimed to engage students meaningfully in assessment rather than giving them tasks simply "for the sake of having an assignment." He offered a choice of tasks that could be assessed using the same criteria. He felt that this was fair because his students could "build on" or "show off" their strengths with "more than one chance or opportunity to be successful" while he maintained consistency in his judgment process. For Kevin this meant that students were "jumping through the same hoops, even though it's a little different . . . everybody in the end is still doing the same thing."

Kevin's recommendations for fair assessment tended to draw on pedagogical knowledge he had developed through experience, but they were also strongly guided by his moral beliefs. For example, his belief in the importance of good work habits surfaced repeatedly. He described an incident involving a student who had not attended class most mornings, but was allowed to pass by the school principal because she had a note from a psychiatrist. Kevin felt that this was *not* "really fair" to the other students who had attended class and “put in” their hours. Kevin frequently referred to student behavior in discussing fair assessment, and he also felt that student factors, such a class size and students’ characteristics, strongly influenced the degree to which assessment could be used for learning.

Lorena: who the learner is

Lorena was a special education teacher, teacher-librarian, and literacy leader who worked at a small public intermediate school (<400 students) in the center of a large urban area. She had been teaching for 14 years. Her education included a Bachelor of Arts (Drama), Bachelor of Education, and additional qualifications in Dramatic Arts, English as a Second Language, Librarianship, Reading, and Special Education, and she was working on becoming a principal. Her school offered, in addition to the regular program, classes for Grade 7 and 8 students with learning, behavioral, developmental, and physical disabilities.

A strong sense of professional care permeated Lorena’s discussion about her practice. Lorena was involved in an action research project on teacher moderation that influenced two of her recommendations for fair assessment. First, she saw a need for consistency in decision-making across classrooms, and second, she emphasized the importance of reflection about assessment. Lorena explained that she and the other Grade 7 English teachers in the project were developing “common goals and common understandings” as they collaborated in marking student writing, and she concluded collegial reflection produced “more fairness and a more level playing field.” Lorena also questioned her own practices, which she attributed to being an “Italian Catholic mother.” She felt that it was crucial for teachers to be aware of their “biases in order to assess fairly” and to examine their teaching when student learning was less than ideal.

Two further recommendations from Lorena related to gathering information about learners and differentiating for individual needs. She gave several examples of situations in students’ lives, such as a parent working several jobs, that could affect learning, and she noted “if we don’t take that into consideration, then how can we be fair?” Lorena explained that information about learners was necessary in order to differentiate appropriately:

> Knowing who the student is, is being fair. It’s like being a parent . . . you’ve got three kids, and one of them has worn out their shoes. Well, in some circles you get everybody a new pair of shoes, but no, it’s only child number two who needs the new pair of shoes, and that’s being fair.

She also stressed that varied information should be considered in assessment decisions, including the learner’s background, student records, input from previous teachers, and reflection on practice. Overall, Lorena’s responses suggested that she felt accountable for her students’ learning, and that she saw nurturing a classroom environment to encourage fair assessment as a teacher’s professional responsibility.

York: as much of the picture as possible

After immigrating from England, York earned a Bachelor of Arts and a Bachelor of Education in Ontario. He had been teaching for 18 years, had additional qualifications in Computers, English, Librarianship, and Special Education, and he had experience scoring provincial assessments. York taught students in Grade 11 and 12 English and Geography in a large public learning center (>2000 students). Located in an inner-ring suburb crisscrossed by highways, York’s school offered multiple programs, such as cooperative education, apprenticeship, special education, and English literacy. The overall aim was to help young adults, many of whom were considered at-risk, to graduate from secondary school.

York’s work environment presented challenges for maintaining the constructive orientation that he considered ideal for fair CA. He associated the use of AfL with fairness, and he recommended developing assessment criteria with students, but he explained that time constraints and student behavior sometimes had a negative impact on the process. What he felt it was most important for teachers to know as much as possible about students and their learning.

> Put yourself in the shoes of the student . . . it’s not about cutting someone a break, or cutting them some slack, it’s about understanding as much of the picture as possible so that when you make a decision it’s based on as many facts as you can.

York was aware of OME policies, and many of his recommendations were consistent with current assessment ideals. However, his concerns tended to focus on the consequences of CA for students. He thought about “student needs” and did what he felt was the “best for the student.” Although this meant that York did not always follow assessment policies, he did engage in critical reflection about his decisions and practices. For example, he described an incident where he had lowered a capable student’s mid-term grade to 1% because of incomplete assignments. After noting the “power and punishment” involved, he explained that he had “no other way” to provoke a change in this student’s work habits. York explicitly recognized the risk of bias, but he maintained that it was crucial for teachers to “know where the student is coming from and who they are as a person.” Ultimately York’s perception of his students’ best interests weighed heavily in his decisions, and he felt this was fairest for them.

Tanar: walk with the student

As the eldest son of Hungarian immigrants with limited English, Tanar became the translator and tutor in his family at a young age. In addition to Bachelors’ degrees in Arts and Education and a Master of Arts (English), and he had additional qualifications in English, Librarianship, and Special Education. He had been teaching for 20 years. He taught Grade 12 Writer’s Craft online for his school board, and he was the teacher-librarian in a mid-sized Catholic secondary school (<1000 students) located in an older suburb. This school offered advanced placement, arts-focused, and gifted programs, and the overall achievement rate on the literacy test was above the provincial average.

Teaching and learning were endlessly interwoven in Tanar’s life, and this influenced his approach to CA. He argued that it had to serve learning to be fair, and he did not limit learning to academics. Tanar valued constructive and respectful assessment interactions, which he felt should be explicitly taught and modeled. An incident that he described as unfair involved being forced, by the principal, to accept an essay from a student who had previously submitted it in another course. Tanar felt that this not only short-changed the student in terms of long-term learning, it was also unfair to other students who had worked harder. This incident and his recommendation that students have the “same opportunity to learn” suggest an overriding belief in equal treatment. However, Tanar also recognized that some circumstances warranted differentiation:

> There are, from time to time, individual cases which require us to think as humane individuals, to adjust assessment . . . and that’s fair too . . . because how can you assess someone who’s lost a mother in a car accident the same way that you can assess someone who’s simply going through the normal ups and down.

When asked if these ideas conflicted, Tanar responded "it's an essential tension ... there's no great solution to this ... a lot of assessment boils down to being judicious." He saw fair assessment as a reflective process that balanced between knowing students and avoiding any bias that might result from that knowledge. While noting that a "personal clash" should not influence assessment decisions, he stressed that teachers should "walk with the student ... put yourself in the student's place." Essentially, Tanar saw fair CA as a human endeavor that was influenced by teachers and students' actions, and thus required reflection to ensure that it supported learning.

Lucy: they've got to earn it

Lucy was on the verge of retirement with 40 years of teaching experience, 21 of which were in Ontario. She began teaching in South Africa after earning two Bachelor of Arts degrees, and she continued her professional education after immigrating to Canada with additional qualifications in Computers, English, English as a Second Language, and Guidance. Lucy taught Grade 10 and 11 students English in congregated gifted classes, and Grade 12 Writer's Craft at a large secondary school (>1000 students) in a suburban neighborhood near a major urban center. Most students spoke languages other than English at home, and the overall achievement rate was above average on the provincial literacy test.

Lucy was an award-winning teacher who held high standards for learning. She worked hard for her students, and she expected a similar effort in return. This approach was reflected in many of her recommendations for fair assessment. First steps that Lucy recommended were for teachers to be aware of the assessment purpose and communicate learning expectations:

> You have to know ... what are you measuring ... you have to know that yourself before you can give out the exercise ... that's critical, and fair means that must be communicated to them because otherwise they'll grope in the dark. If they don't know what your expectations are, you can't be fair.

When asked what was most important for fairness, Lucy identified consistency, which she defined as marking "to one standard and one standard only." While she stressed this repeatedly, she did not see differentiation as unfair, and she believed that "making allowances" was congruent with her teaching philosophy. Lucy gave examples of adapting an assessment task to meaningfully engage gifted students and modifying another for a student with Asperger's syndrome to illustrate the notion that assessment is fairer for students if it provides an appropriate learning challenge.

Lucy's acceptance of differentiation came with a caveat. She expressed a strong moral belief in the merit of hard work, and she felt that students did not "gain anything" when teachers devalued learning by "giving away marks." Lucy declared that students should "really earn" their grades, and expecting less was a "disservice" to them.

> Fairness for me is knowing that the student has done the job honestly and with a maximum effort in order for me to ... give them the advantages that they might not have had in the first place ... So it really is based on my sense of how honest they are, the sincerity of their need, and their performance.

In sum, Lucy offered several interpretations of fairness, but her belief in the importance of hard work was dominant, both in what she expected from and provided for her students.

Amada: never settle for mediocrity

Amada was a retired teacher and school leader who continued to tutor students and teach on-call at an adult learning center. She had been teaching for 40 years, 38 of which were in Ontario, and she was passionate about sharing her knowledge. Amada explained that she had learned to value education as a means of social mobility while attending a private high school in Jamaica. After immigrating to Canada, she steadily pursued higher education, which resulted in a Bachelor of Arts, Master of Arts (Spanish), Master of Education, and additional qualifications in Adult Education, English, English as a Second Language, French as a Second Language, Principalship, Reading, and Spanish. Amada lived and worked in a densely populated and culturally diverse area that had been subsumed through the expansion of a major urban area. Most of the students she worked with were born outside of Canada and aimed to improve their English writing skills.

Amada felt that there were many ways of being fair in CA. For example, she said it was "not fair to constantly mark for everything" because criteria should be feasible for students. She also thought that it was important to give honest and accurate feedback, while being "democratic" in distributing attention and praise among students.

A belief that reverberated through Amada's responses was that fairness depends on a teacher's knowledge and ability. In this she included content knowledge, pedagogical knowledge specifically about CA, general knowledge of students, an understanding of particular students, and a teacher's ability to respond constructively to student learning. She explained that to design fair assessment criteria and tasks, a teacher needs a "high level of mastery of subject matter" and a "certain modicum of common sense in dealing with young people" in order to "balance ... knowledge of where the student is, what the student is capable of doing, and what they can realistically demand." Amada also saw a critical role for teachers' knowledge in a second area of concern: orienting assessment to constructively support high expectations for learning. Amada thought of fair CA as an ongoing interaction that focused on students' learning, and she emphasized the need for teachers to be responsive in this process.

> As the student grows, then you move the yardstick a little bit more ... the teacher should never be contented with whatever effort that is given by the kid. Praise him, praise her for whatever she's done, but always keep saying ... there's a little more that you can do ... never settle for mediocrity.

From Amada's perspective, the fairness of CA rested squarely in teachers' hands; it depended not only on teachers' knowledge, but also on their ability to respond constructively to learning.

Multi-faceted fairness in classroom assessment

Commonalities and differences were seen in the participants' recommendations for fair CA. They drew to varying degrees on their experiences as learners and teachers, their theories about learning and learners, and their content knowledge. They were also influenced by their moral beliefs, their knowledge of Ontario's educational policies and politics, and the particulars of their work environments. Looking across the cases, five needs for fairness are seen: opportunities for students, transparency in assessment procedures, critical reflection about assessment practices, constructive classroom environment, and a balance between equality and equity. These are discussed in relation to existing literature highlighting how they contribute to thinking about fairness as a multifaceted quality in CA.

Students' opportunities

The participants repeatedly and consistently stated in their responses to the vignettes that students should be given ample opportunity to learn. Unlike some teachers in empirical studies (e.g., Eggen, 2004; Gummer, 2000; Zoeckler, 2005), they were less concerned with all students having the same opportunities, and they tended to focus on learning needs. From a technical perspective, opportunity to lean is the alignment between instruction and assessment (McMillan, 2011; Russell & Airasian, 2012), but it is also a social justice issue stemming from the differences in students' previous learning opportunities (Drummond, 2003; Gee, 2003; Weeden, Winter, & Broadfoot, 2002). The participants' poignant stories about their own learning experiences and their concern about their students' welfare suggested that they were well aware of the latter. For example, one rationalized that boosting grades for 'at-risk' students was fair because they had been dealt fewer opportunities than typical Ontarians. However, none of the participants mentioned using AfL to minimize students' disparate opportunities, which has been suggested in CA literature (Camilli, 2006; Crossouard, 2011; Poehner, 2011; Pryor & Crossouard, 2008; Wormeli, 2006). At that point, they did not explicitly recognize the role of AfL in their advocacy for students, possibly because the province was in the early phases of shifting to AfL practices from a grading culture.

Most participants were concerned about the accuracy of grades, and their impact on students' future opportunities, specifically in terms of students' obtaining scholarships, succeeding in university, or gaining employment. Morgan and Watson (2002) argue that the accuracy of teachers' assessments can lead to unfairness because they have the potential to limit students' future opportunities. Like other teachers in Ontario, the participants were confident about their assessment practices (Tierney, Simon, & Charland, 2011), and they did not dwell on their own inaccuracies. Their concerns centered on the relationship between students' work habits, final grades, and future opportunities. Two felt that intervention by principals had resulted in grades that were unfair because they over-represented student effort, and two argued that easy grading was unfair because it did not realistically reflect achievement. All were concerned that about preparing students for success in future opportunities. As such, their recognition of a fairness issue focused less on technical accuracy and more on the responsibility of teachers within the broader social purpose of schooling.

When the participants went beyond responding to the vignettes and described their own practices, they focused on students' opportunities to *demonstrate* learning. Several suggested that to be fair, assessment decisions should be well informed and provide a complete picture of student learning. The use of multiple assessments is recommended for fairness in assessment theory (e.g., Camilli, 2006), and it is generally accepted among teachers (CMEC, 2002; Green et al., 2007). This facet of fairness is akin to reliability conceptualized as a sufficiency of information (Smith, 2003). Many types of assessment-based decisions are fairer when multiple scores or sources are used. However, one participant stressed that to be fair, teachers should continuously modify their expectations as they come to know and respond to students' learning. This is consistent with the observation that information gathered in ongoing classroom interactions is necessarily incomplete (Bulterman-Bos, Terwel, Verloop, & Wardekker, 2002; Morgan & Watson, 2002). In this light, the evidence needed for fairness differs with purpose. For fairness in AfL, a real-time stream of frequently refreshed information about learning is needed rather than a summative snap shot.

Two participants in this study tailored performance assessments to meaningfully engage or challenge their students, which they felt was fairer than a one-size-fits-all approach. Opportunity to demonstrate learning is enhanced when teachers hold high expectations (Campbell, 2003; Wiggins, 1993), and they plan activities for substantive learning (Brighton, 2003). Students' opportunities should be varied to allow diverse students to show what they know in different ways (Heritage, 2013; Suskie, 2002). From a moral perspective, learning cannot be fully acknowledged or valued with insufficient opportunity for its demonstration (Buzzelli & Johnston, 2002). From a technical perspective, student engagement is important because assessment tasks require the "willing cooperation of the person whose competency is being measured" (Badger, 1999, p. 68). This is also a matter of equity as tasks that are designed to engage diverse students are more likely to value different kinds of knowledge or manifestations of learning in the classroom (Cowie, 2014). Given learner diversity, the complexity of capturing learning, the time needed for performance assessments, and the brevity of reporting periods in many educational systems, these two participants make an important point. The fairest and most effective approach in these circumstances is to assume that all CAs should serve learning, even when their main purpose is AofL.

Transparent communication

The participants unanimously recommended that learning expectations and assessment criteria be shared with students, and they all emphasized that this should be done the beginning of an assignment. Assessment specialists have long recommended clear communication with students about assessment (e.g., Andrade, 2005; Guskey & Jung, 2009; JAC, 1993; JCSEE, 2003; McMillan, 2011; Sadler, 1989; Shepard, 2000; Speck, 1998; Suskie, 2002), and research with teachers suggests wide agreement (e.g., Brookhart, 1994; Green et al., 2007; James & Pedder, 2006a; Ploegh, Tillema, & Segers, 2009; Tierney et al., 2011; Zoeckler, 2005). Sharing expectations and criteria with students also figures prominently in guidelines for AfL (ARG, 2002; Stiggins & Chappuis, 2005). Thus, the participants' phronesis about transparency was congruent with a principle that is well-accepted in the CA community.

A difference of opinion emerged in relation to rubrics. Explicit criteria support learning by providing students with a framework for metacognition and encouraging learning-oriented dialog about the desirable qualities of a task (Andrade, 2005; Shepard, 2006). However, they can also normalize the knowledge gained and displayed by students, and restrict that which is recognized and valued by teachers (Crossouard, 2011; Morgan & Watson, 2002). This has made the use of rubrics controversial for teaching and assessing writing (e.g., Badger, 1999; Newkirk, 2004; Wilson, 2006). Some participants suggested strategies, such as using exemplars and student-generated criteria, to ensure that rubrics enhance fairness without constraining learning. While evidence regarding the effectiveness of these strategies is not widely established, it is emerging (e.g., Andrade, 2013; Andrade, Du, & Wang, 2008). The participants differed on two further points: first in their ideas about whether it was fairer to give students feedback during or after a writing activity, and second in their beliefs about the importance of correcting grammatical errors. Their disagreement reflects a controversy among English teachers about what should be valued in teaching and assessing writing, which emanates from a broader shift in social ideals about the purposes and goals of education.

Constructive classroom environment

The participants all recommended strategies for teachers to proactively nurture a constructive learning environment as the setting for fair assessment. This idea is not seen in the interpretations of fairness for testing (AERA, APA, & NCME, 1999), nor is it explicit in the operational definitions for CA (Camilli, 2006; McMillan, 2011; Russell & Airasian, 2012; Suskie, 2002; Volante, 2006). However, it is not entirely surprising given the increasing recognition of the social nature of learning.

Following an early description of the CA environment that focused on teachers' practices (Stiggins & Bridgeford, 1985), it became apparent that a better understanding of the dynamics of assessment interactions (e.g., questioning, feedback, peer assessment) was needed for assessment to genuinely support learning (Torrance & Pryor, 1998). The importance of trust and respect in the learning process has since been highlighted (Cowie, 2005; Shepard, 2006), along with the complexity of renegotiating power and identity in CA (Pryor & Crossouard, 2008). Research on the relationship between the assessment environment and learning is ongoing. For example, Dorman, Fisher, and Waldrip (2006) found that students' perceptions of assessment tasks influence their sense of self-efficacy and their attitudes about learning. Birenbaum, Kimron, and Shilton (2011) also found an association between the classroom culture, the school learning community, and AfL. At this point, the idea that a constructive environment has benefits for leaning is well accepted. It is emphasized by assessment specialists (e.g., Russell & Airasian, 2012) and endorsed by the broader educational community (JCSEE, 2003). What is not as clear is the relationship between the fairness of CA and the learning environment.

The phronesis offered by the participants suggests a symbiotic relationship between fair assessment and AfL. They are not the same, but they can be mutually supportive. Several participants stressed that actively promoting respectful interactions in CA is a teacher's professional responsibility. In Ontario, four ethical standards for teaching focus on respect, trust, care and integrity, but they offer little guidance for CA (OCT, 2010). Cowie (2014) proposed an ethical principle specifically for AfL, where "harm is minimized through respectful relationships" (np) in the classroom. Research on students' perceptions of fairness indicates that trust and respect in assessment interactions is widely valued (Gordon & Fay, 2010; Tata, 2005; Wendorf & Alexander, 2005). Many of the participants' recommendations focused on trust and respect, and they took the public nature of CA into account. One participant recommended teaching students how to give constructive feedback directly and through modeling in order to avoid the problem of "poisoning" (Pryor & Lubisi, 2002, p. 682) relationships during peer assessment. While modeling is not commonly associated with fairness in assessment literature, it is included in one operational definition (Camilli, 2006). This could be developed as a form of "metacontextual reflection" (Pryor & Crossouard, 2008, p. 16) that includes fairness in framing AfL. Essentially, reflection is situated within, but also influences the assessment environment. As Brookhart (2003) observed, the outcomes of CA "fold back into and become part of the classroom environment" (p. 7). With this in mind, it can be said that a constructive environment is needed for fair AfL interactions, while fairness in AfL interactions enables the environment to genuinely support learning.

Reflective interaction

Two participants discussed the importance of teachers' critical reflection for fair assessment, and thoughts about power, biases, and the purpose of assessment were sprinkled through their responses. The need for reflection to encourage fairer assessment has been repeatedly noted in assessment literature. Topics for reflection include children's rights (Drummond, 2003), power and control (Cowie, 2014; Gipps, 1999; Schendel and O'Neill, 1999), personal values (Whittington, 1999), biases and stereotypes (Andrade, 2005; Camilli, 2006; McMillan, 2011; Popham, 2008; Russell & Airasian, 2012; Volante, 2006; Weeden et al., 2002), the intersection of gender and achievement (Elwood, 2006b), assessment outcomes (Suskie, 2002), and assessment purposes (Delandshere, 2001; Earl & Katz, 2006; Speck, 1998). Essentially, these calls for reflection aim to disrupt assumptions and help improve how learning is assessed. While they may generate discussion and further research, it is not evident that they inform reflective practice. At present, the degree to which critical reflection is intentionally used in teaching practice as strategy for fairer assessment is not evident.

Previous research has found that teachers have difficulty describing their assessment practices (Dixon & Williams, 2003) and explaining their decisions (McMillan, 2003). Teachers' vocabularies appear to be insufficient in some contexts for critical discussion about the complexities of CA (Black, Harrison, Hodgen, Marshall, & Serret, 2010; Crossouard, 2011). The participants in this study had no difficulty describing their assessment practices, possibly due to their purposeful selection. They were all able to elucidate their understanding of fairness as a quality of CA, with the two most experienced teachers being especially fluid in articulating their phronesis. Nonetheless, all of their words were halting when they tried to identify the types of knowledge and beliefs they had drawn on for their recommendations. This may be because phronesis is both dynamic and situated. As Dunne (1993) notes, we can "never quite catch up" in understanding our thoughts because we are "always already beholden to assumptions, antecedent interests and tacit procedures which are not themselves known" (p. 357). Responding in writing to the vignettes before the interviews facilitated discussion because it accorded time for reflection. Explaining responses out loud sparked additional reflection, and some thoughts that might otherwise have remained dormant were then spoken. It seems that even though phronesis is individual and rooted in the particulars of practice, discussing it with others helps it flourish.

Critical reflection about assessment purposes may be particularly important for teachers in educational contexts where policies emphasize AfL while requiring regular AofL. In England (Black, Harrison, Lee, & Marshall, 2003; Torrance & Pryor, 2001) and the United States (Brookhart, 2001) classroom-based research has shown that the theoretical distinction between assessment purposes blurs in practice. It is possible, as the participants suggested here, for students and teachers to use AofL results fairly and effectively for learning. However, reverse usage threatens fairness. Using information gathered during teaching and learning to report on achievement is problematic, especially if students are encouraged to disclose misconceptions or weaknesses under the guise of AfL. Reflection about fairness for different assessment purposes also needs to be encouraged because the concerns teachers express tend to relate solely to AofL (e.g., Eggen, 2004; Zoeckler, 2005), and fairness issues in AfL are overlooked. In this study, where the vignettes encouraged reflection about AfL, the issues highlighted by the participants based on their own practices still involved fairness in AofL. Thinking about fairness in testing and AofL is important, but its dominance overshadows the urgent need for reflection about fairness in the dynamics of AfL.

A final point on reflection about fairness relates to the people involved. One participant felt that conversations about assessment with colleagues could lead to greater fairness. Studies on professional development highlight the benefits of collegial discussion, but also note potential drawbacks (e.g., Lovett & Gilmore, 2003; James & Pedder, 2006b). Some of the teachers' narratives in Campbell's (2003) study illustrate the dark side of collegiality where fear of repercussion engenders silence. The benefit of assessment conversations should not be assumed, and the social contexts that affect group discussion must be taken into account (Hollander, 2004). Furthermore, this should include everyone involved in assessment conversations. Wyatt-Smith, Klenowski, and Gunn (2010) analyzed dialog among teachers marking in a large-scale assessment program, and they concluded that "active listening to what transpires in moderation" (p. 73) is needed in contexts where teacher judgment plays a central role. Considered in tandem with the participants' ideas about actively promoting an environment for fair assessment, the same argument exists for listening more closely to what is said and implied during assessment conversations in classrooms, including interactions between peers. Although engaging students in CA is advantageous for learning (Black & Harrison, 2001; McDonald & Boud, 2003), fairness issues arise when power dynamics (Brookfield, 2001; Dann, 2002) and the quality of students' interactions (Pryor & Lubisi, 2002) are not considered. When learners are involved in CA, they should be encouraged to reflect about the purpose of their judgments and the biases involved.

Equality in the classroom

The idea that learners should be treated equally ran through the participants' recommendations. Most felt that everyone in a class should complete the same amount of work. Several expressed concerns about inconsistencies in grading, and they felt that special treatment was unfair unless clearly warranted by the students' circumstances. Many teachers believe that consistency in CA is needed for fairness (e.g., Allen & Lambating, 2001; Campbell, 2003; Eggen, 2004; Ryan, 2000; Szpyrka, 2001; Yip & Cheung, 2005; Yung, 2001; Zoeckler, 2005). It is indeed important when test scores or assessment results are used to compare or rank students, and guidelines for CA state that students should have the *same* information about an assessment and the *same* opportunities to demonstrate learning (JAC, 1993; JCSEE, 2003). However, this was not the underlying reason for the participants' emphasis on equal treatment. Their phronesis was permeated by strong moral beliefs about the value of effort and their professional obligation to encourage good work habits. This is reflective of a standard of practice in Ontario, which directs teachers to "facilitate the development of students as contributing citizens of Canadian society" (OCT, 2010). More generally, work-world pressures have produced "universal agreement" on one function of schooling, namely that students should learn a strong "work ethic" (Leibowitz, 2000, p. 62). With this in mind, the participants concerns about equal treatment in AofL were aligned with a value that is strongly held in the broader social context.

Students tend to associate fairness with equal treatment in the classroom (Bursuck, Munk, & Olson, 1999; Dalbert et al., 2007; Duffield & Spencer, 2002; Robinson, 2002; Smith & Gorard, 2006), and teachers echo their concerns (Brighton, 2003; Brookhart, 1993; Torrance & Pryor, 2001). Although each participant described assessments where they felt differentiation was fair, some aimed to appear fair by minimizing differences in how they treated students. What seems at the heart of their concern was the *appearance* of equality, rather than equality itself. The equation of fairness with equal treatment in the educational community may change over time with statements in assessment policy documents and teaching materials such as "fairness is not sameness" (OME, 2007, p.28) and "fair isn't always equal" (Wormeli, 2006, p. 6). In the meantime, for differentiation to be accepted among peers and useful for learning, teachers will need to keep students' perceptions about equality in mind, particularly when the learning environment is taking shape at the beginning of a course or year.

Equitable treatment

In contrast to their ideas about equal treatment, the participants voiced no concerns about inconsistency in students' learning opportunities. This may relate to some degree to the timing of assessment events described in the vignettes. Differentiation occurring early in a unit may have been perceived as fair because it is associated with teaching and learning, rather than evaluative judgments. This would be consistent with some of the teachers in Brighton's (2003) study who were enthusiastic about individualizing instruction, but found differentiation in grading problematic. However, responses to a vignette that described a teacher giving feedback suggest that timing may not be the only factor in the tension between consistency and differentiation. None of the participants were concerned about students receiving different amounts of feedback, but there was disagreement about the tone. One recommended that feedback be as positive as possible, while the others felt it should refer to students' strengths and weaknesses. Wiggins (1993) argued that fair assessment "properly balances challenge and possible success" (p. 174). Novice teachers sometimes worry that honest feedback will damage students' self-esteem (e.g., Graham, 2005), but indiscriminately positive feedback fails to recognize individual achievement (Buzzelli & Johnston, 2002; Campbell, 2003; Jackson, Boostrom, & Hansen, 1993). The two most experienced participants felt strongly about the need for honest and balanced feedback, and they emphasized that anything less was unfair because it would have negative consequences for students' learning in the long term. This suggests another difference for fairness related to assessment purposes. AfL is often student-referenced, and thus inherently varied. Equitable differentiation in feedback during AfL is needed to maximize learning before AofL.

The idea that assessment should be equitable emanates from three areas: (a) the view that varied knowledge brought to classrooms by culturally diverse learners should be recognized as a wealth (e.g., Cowie, 2014; Jiménez, 2004/2005; Tierney, 1998/2005), (b) theories that integrate teaching, learning, and assessing (e.g., Moon, 2005; Poehner, 2011; Tomlinson, 2005), and (c) special education guidelines (e.g., Guskey & Jung, 2009). The latter have increasingly been adopted in assessment policies and teaching practice. For example, the OME (2010) discusses modifications and accommodations for specific groups of learners extensively, and the teachers surveyed by Green and colleagues (2007) strongly agreed (94%) that a student with a learning disability should receive an accommodation during a classroom test. The participants associated accommodations with learning difficulties, and they felt that differentiated assessment was fair for students who faced some type of challenge. Their stance in this regard was similar to the idea of "pulling for students" that McMillan (2003, p. 37) observed in teachers' assessment practices. However, when issues arose with students who were perceived as academically capable, the participants were less likely to recommend differentiated assessment as a fair alternative. Furthermore, two participants felt that student effort had to be considered for *any* form of differentiation. Several studies have found that teachers consider effort and are more lenient in grading academically weaker students (Resh, 2009; Zoeckler, 2005). Brookhart (1993) called this phenomena the "double standard of just deserts" (p. 140). Resh (2009) used the concept of "deservedness" (p. 317) to describe how teachers apply varying combinations of achievement, effort and need in determining students' grades. A similar idea, that students' marks should reflect what they deserve, also surfaced in the teacher moderations sessions studied by Wyatt-Smith et al. (2010). The problem with this for fairness, in both AfL and AofL, lies in the perception of deservedness. Students display effort and need differently, and teachers' observations are influenced by time, class size, their theories about teaching, learning and assessing, and any biases or stereotypical beliefs they might have about learners (Brighton, 2003; Bulterman-Bos et al., 2002; Watson, 1999). In the classroom, this can result in compounded perceptions that influence the quality of assessment interactions and consequent opportunities. Beyond the classroom, it can mean that the distribution of future opportunity based on students' grades is unfair for many students.

Conclusion

Fairness in CA is a complex quality that relies heavily on teachers' professional judgment. Six experienced and knowledgeable teachers shared their phronesis about fairness in CA, which I considered in relation to existing literature. I weighed several potential limitations in this study's design, two of which relate to self-reported information. First, self-reports are not necessarily congruent with action (Wilks, 2004), meaning here that a participant might give a recommendation without actually practicing it. When I undertook this project it was already evident that teachers were concerned about fairness in CA (e.g., Yung, 2001), but explicit and coherent theory to guide practice was lacking. As such, I was more interested developing CA theory that is grounded in practice and addresses teachers' concerns than I was in further observation of practice. A second limitation is that self-reports have the potential to instigate socially desirable responding (Holtgraves, 2004). Vignettes help reduce this by directing initial attention away from participants (i.e., on a salient story), and giving them control over when and to what degree they share personal experiences (Barter & Renold, 2000). A strategy that encourages disclosure in vignette research is to have participants "adopt the role of consultants" (Hughes & Huby, 2004), which I did by articulating that I valued their thinking and by asking them to provide recommendations for other teachers and researchers. Nonetheless, the participants were responding with sufficient knowledge of assessment policies in Ontario to permit politically correct responses. Many of their recommendations did align with current assessment ideals, but they also described situations which were less than perfect in terms of fairness. In critical pragmatism, beliefs and opinions cannot be separated as 'true' from the time and place in which they are situated. Thus the participants' understanding of their context, their moral beliefs, their theoretical and practical knowledge, and their varied experiences were accepted as inherent to their thinking about a complex concept and social issue.

My intent in undertaking this study was neither to prescribe a fixed set of steps for all teachers to follow, nor to suggest that fairness is manifested consistently across all classrooms. Essentially, I aimed to contribute knowledge from a particular context to better understand a universal concept (Erickson, 1986; Merriam, 1998). Viewed from a nomothetic perspective, this multi-case study's lack of statistical generalizability is a limitation (Williams, 2004). While it is often assumed that generalization from case studies is not possible, there are many examples of general knowledge having been built from particular situations or events, and different types of generalizations can be made from qualitative research (Eisenhart, 2009; Flyvbjerg, 2006). Two types are possible from this work. The first is generalization to similar contexts, which Guba and Lincoln (1994) referred to as "transferability" (p. 114). A clear description of the context and sampling supports transferability (Eisenhart, 2009; Jensen, 2008). The results of this study could shed light on fairness issues in similar educational systems, especially where CA practices are shifting toward a greater acceptance of AfL. The second type is "theoretical generalization" (Eisenhart, 2009, p. 60), which is supported by my purposeful selection of participants based on their likelihood of contributing phronesis to the reconceptualization of fairness. Theoretical generalization occurs when particular cases help nuance and update evolving theory. Thus, while it is not my intent to generalize in the traditional quantitative sense, this multi-case study does contribute valuable information for thinking about fairness in CA.

Similarities do exist between fairness in standardized testing and CA. Fairness is a matter of degree in both, and not a dichotomous characteristic. It also relies, in both, on a combination of technical skill, thoughtful planning, and a review process. Finally, a similar social-moral imperative is seen in the emphasis on merit in the history of testing and in teachers' beliefs about CA. However, standardized tests and CA differ in how they capture learning. While one is a static measure, the other requires teachers to continuously update information about students' learning. Thus, fairness can never be fully established in CA; it is an ongoing endeavor that fluctuates in intensity with cycles of teaching, learning and assessing. Fairness in AfL is particularly complex because it is a quality of human interactions in perpetually evolving circumstances. A posteriori evidence determined through investigation or argument is helpful for fairness from a theoretical perspective, but it is insufficient for the immediacy of classroom practice, especially for the learners involved.

I have suggested multiple avenues for research on fairness in CA elsewhere (Tierney, 2008, 2013). However, there are some that merit mention here. To date I have focused on highlighting the *from* and *to* in the process of rethinking fairness, and it would be beneficial to look beyond measurement and CA theory to consider how fairness and related concepts are understood in other fields. For example, findings from research on distributive, procedural and interpersonal justice in organizations could expand current understanding of the relationship between fair assessment and the learning environment. Promising work has been done in this vein on grading (e.g., Resh, 2009; Tata, 2005), and it should also be pursued for AfL. Another thread that should be pursued relates to teacher education. Two research questions in particular arise from this work. First, to what degree in teacher education programs is CA discussed as an inquiry process that requires self-awareness and reflection, rather than as a set of tools with universal application? Second, how are novice teachers oriented to think about their assessment interactions with diverse learners? This leads to a third area that should also be investigated. Given that classrooms vary considerably, comparative research would be useful for understanding the degree to which different aspects of fairness are valued or pertinent in different educational contexts. To better understand how assessment might serve diverse students, inquiry into stakeholders' interpretations of fairness is needed at all levels (i.e., primary to higher education) and in different cultural contexts. Assessment cannot be assumed by the educational community to be universally fair simply because it aims to support learning. Ongoing research and reflection about fairness issues in CA are needed to encourage and sustain fairer assessment practices for diverse learners.

References

Abrams, L. M., Pedulla, J. J., & Madaus, G. F. (2003). Views from the classroom: Teachers' opinions of statewide testing programs. *Theory into Practice, 42*(1), 18.

Allen, J. D., & Lambating, J. (2001). *Validity and reliability in assessment and grading: Perspectives of preservice and inservice teachers and teacher education professors*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education, (1966). *Standards for educational and psychological tests*. Washington, DC: American Educational Research Association.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education, (1974). *Standards for educational and psychological tests*. Washington, DC: American Educational Research Association.

Andrade, H. G. (2005). Teaching with rubrics: The good, the bad, and the ugly. *College Teaching, 53*(1), 27–30.

Andrade, H. G. (2013). Classroom assessment in the context of learning theory and research. In J. H. McMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 17–34). Thousand Oaks, CA: Sage Publications.

Andrade, H. L., Du, Y., & Wang, X. (2008). Putting rubrics to the test: The effect of a model, criteria generation, and rubric-referenced self-assessment on elementary school students' writing. *Educational Measurement: Issues and Practice, 27*(3), 3–13.

Aristotle. (trans. 1925/1975). *The Nicomachean ethics of Aristotle*. London: Oxford University Press.

Assessment Reform Group. (1999). *Assessment for learning: Beyond the black box*. Cambridge, UK: Cambridge University.

Assessment Reform Group. (2002). *Assessment for learning: 10 principles*. Retrieved from http://www.assessment-reform-group.org/publications.html.

Badger, E. (1999). Finding one's voice: A model for more equitable assessment. In A. L. Nettles & M. T. Nettles (Eds.), *Measuring up: Challenges minorities face in educational assessment* (pp. 53–69). Norwell, MA: Kluwer Academic Publishers.

Baartman, L. K., Bastiaens, T. J., Kirschner, P. A., & van der Vleuten, C. P. M. (2006). The wheel of competency assessment: Presenting quality criteria for competency assessment programs. *Studies in Educational Evaluation, 32*, 153–170.

Baartman, L. K., Bastiaens, T. J., Kirschner, P. A., & van der Vleuten, C. P. M. (2007). Teachers' opinions on quality criteria for competency assessment programs. *Teaching and Teacher Education, 23*, 857–867.

Barter, C., & Renold, E. (2000). I wanna tell you a story': Exploring the application of vignettes in qualitative research with children and young people. *International Journal of Social Research Methodology, 3*(4), 307–323.

Bernstein, R. J. (1985). *Beyond objectivism and relativism: Science, hermeneutics and praxis*. Philadelphia, PA: University of Pennsylvania Press.

Biesta, G. J. (2007). Why "what works" won't work: Evidence-based practice and the democratic deficit in educational research. *Educational Theory, 57*(1), 1–22.

Biesta, G., & Burbules, N. C. (2003). *Pragmatism and educational research*. Lanham, MD: Rowman & Littlefield Publishers Inc.

Birenbaum, M., Kimron, H., & Shilton, H. (2011). Nested contexts that shape assessment for learning: School-based professional learning community and classroom culture. *Studies in Educational Evaluation, 37*, 35–48.

Black, P., & Harrison, C. (2001). Self- and peer-assessment and taking responsibility: The science student's role in formative assessment. *School Science Review, 83*(302), 43–49.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. Maidenhead, England: Open University Press.

Black, P., Harrison, C., Hodgen, J., Marshall, B., & Serret, N. (2010). Validity in teachers' summative assessments. *Assessment in Education: Principles, Policy & Practice, 17*(2), 215–232.

Black, P., & Wiliam, D. (2006). The reliability of assessments. In J. Gardner (Ed.), *Assessment and learning* (pp. 119–131). London: Sage Publications.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74.

Bonner, S. M. (2013). Validity in classroom assessment: Purposes, properties, and principles. In J. H. McMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 87–106). Thousand Oaks, CA: Sage Publications.

Brighton, C. M. (2003). The effects of middle school teachers' beliefs on classroom practices. *Journal for the Education of the Gifted, 27*(2/3), 177–206.

Brookfield, S. (2001). Unmasking power: Foucault and adult learning. *Canadian Journal for Studies in Adult Education, 15*(1), 1–23.

Brookhart, S. M. (1993). Teachers' grading practices: Meaning and values. *Journal of Educational Measurement, 30*(2), 123–142.

Brookhart, S. M. (1994). Teachers' grading: Practice and theory. *Applied Measurement in Education, 7*(4), 279–301.

Brookhart, S. M. (2001). Successful students' formative and summative uses of assessment information. *Assessment in Education: Principles, Policy & Practice, 8*(2), 153–169.

Brookhart, S. M. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues and Practice, 22*(4), 5–12.

Brookhart, S. M. (2004). Classroom assessment: Tensions and intersections in theory and practice. *Teachers College Record, 106*(3), 429–458.

Brookhart, S. M. (2013). Classroom assessment in the context of motivation theory and research. In J. McMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 35–54). Thousand Oaks, CA: Sage Publications.

Bulterman-Bos, J., Terwel, J., Verloop, N., & Wardekker, W. (2002). Observation in teaching: Toward a practice of objectivity. *Teachers College Record, 104*(6), 1069–1100.

Bursuck, W. D., Munk, D. D., & Olson, M. (1999). The fairness of report card grading adaptations: What do students with and without learning disabilities think? *Remedial and Special Education, 20*(2), 84–92.

Buzzelli, C. A., & Johnston, B. (2002). *The moral dimensions of teaching: Language, power, and culture in classroom interaction*. New York: RoutledgeFalmer.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 221–256). Westport, CT: Praeger Publishers.

Campbell, E. (2003). *The ethical teacher*. Maidenhead, England: Open University Press.

Clandinin, D. J., & Connelly, F. M. (1996). Teachers' professional knowledge landscapes: Teacher stories—stories of teachers—school stories—stories of schools. *Educational Researcher, 25*, 24–30.

Clarke, M. M., Madaus, G. F., Horn, C. L., & Ramos, M. A. (2000). Retrospective on educational testing and assessment in the 20th century. *Journal of Curriculum Studies, 32*(2), 159–181.

Cole, N. S., & Zieky, M. J. (2001). The new faces of fairness. *Journal of Educational Measurement, 38*(4), 369–382.

Connolly, M. R. (2005). *Exploring cases of practical wisdom (phronesis) in postsecondary teaching*. (Ph.D. dissertation) USA: Indiana University. UMI #3167788.

Council of Ministers of Education Canada. (2002). *SAIP Context Data: Teacher Questionnaire/SAIP 2002 – Writing*.

Cowie, B. (2005). Student commentary on classroom assessment in science: A sociocultural interpretation. *International Journal of Science Education, 27*(2), 199–214.

Cowie, B. (2014). Equity, ethics and engagement: Principles for quality formative assessment in primary science classrooms. In C. Milne, K. G. Tobin, & D. Degenero (Eds.), *Sociocultural studies and implications for science education: The experiential and the virtual*. Dortrecht, The Netherlands: Springer.

Cowie, B., & Bell, B. (1999). A model of formative assessment in science education. *Assessment in Education: Principles, Policy & Practice, 6*(1), 101–116.

Crossouard, B. (2011). Using formative assessment to support complex learning in conditions of social adversity. *Assessment in Education: Principles, Policy & Practice, 18*(1), 59–72.

Dalbert, C., Schneidewind, U., & Saalbach, A. (2007). Justice judgments concerning grading in school. *Contemporary Educational Psychology, 32*(3), 420–433.

Dann, R. (2002). *Promoting assessment as learning: Improving the learning process*. London: RoutledgeFalmer.

Delandshere, G. (2001). Implicit theories, unexamined assumptions and the status quo of educational assessment. *Assessment in Education: Principles, Policy & Practice, 8*(2), 113–133.

Dixon, H., & Williams, R. (2003). Teachers' understanding and use of formative assessment in literacy learning. *New Zealand Annual Review of Education*, 1–12,. Retrieved from http://assessment.tki.org.nz/Research/Research.

Doherty, G. D. (2008). On quality in education. *Quality Assurance in Education, 16*(3), 255–265.

Dottori, R. (2009). The concept of phronesis by Aristotle and the beginning of hermeneutic philosophy. *Etica & Politica,* Ethics & Politics*11*(1), 301–310.

Erickson, F. (1986). Qualitative methods in research on teaching. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed.). New York: MacMillan Publishing Company.

Dierick, S., & Dochy, F. (2001). New lines in edumetrics: New forms of assessment lead to new assessment criteria. *Studies in Educational Evaluation, 27*, 307–329.

Dorman, J. P., Fisher, D. L., & Waldrip, B. G. (2006). Classroom environment, students' perceptions of assessment, academic efficacy and attitude to science: A LISREL analysis. In D. L. Fisher & M. S. Khine (Eds.), *Contemporary approaches to research on learning environments: Worldviews* (pp. 1–28). Hackensack, NJ: World Scientific.

Drummond, M. J. (2003). *Assessing children's learning* (2nd ed.). London: David Fulton Publishers Ltd.

Duffield, K. E., & Spencer, J. A. (2002). A survey of medical students' views about the purposes and fairness of assessment. *Medical Education, 36*(9), 879–886.

Dunne, J. (1993). *Back to the rough ground: Phronesis and techne in modern philosophy and in Aristotle*. Notre Dame, IN: University of Notre Dame Press.

Earl, L., & Katz, S. (2006). *Rethinking classroom assessment with purpose in mind: Assessment for learning, assessment as learning, assessment of learning.* Governments of Alberta, British Columbia, Manitoba, Northwest Territories, Nunavut, Saskatchewan, and Yukon Territory: Western and Northern Canadian Protocol for Collaboration in Education. Retrieved from http://www.wncp.ca/english/subjectarea/classassessment.aspx.

Eisenhart, M. (2009). Generalization for qualitative inquiry. In K. Ercikan W.-M. Roth (Eds.), *Generalizing from educational research: Beyond qualitative and quantitative polarization* (pp. 51–66). New York: Routledge/Taylor & Francis.

Eggen, A. B. (2004). *Alfa and Omega in student assessment: Exploring identities of secondary school science teachers.* (Ph.D. dissertation) Norway: Universitetet i Oslo. Retrieved from www.ils.uio.no/forskning/pdh-drgrad/.../AstridEggenAvhandling1.pdf.

Elwood, J. (2006a). Formative assessment: Possibilities, boundaries and limitations. *Assessment in Education: Principles, Policy & Practice, 13*(2), 215–232.

Elwood, J. (2006b). Gender issues in testing and assessment. In C. Skelton, B. Francis, & L. Smulyan (Eds.), *The SAGE handbook of gender and education* (pp. 262–278). Thousand Oaks, CA: Sage Publications.

Erzberger, C., & Kelle, U. (2003). Making inferences in mixed methods: The rules of integration. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 457–488). Thousand Oaks, CA: Sage Publications.

Fenstermacher, G. D. (1994). The knower and the known: The nature of knowledge in research on teaching. *Review of Research in Education, 20*, 3–56.

Finklestein, I. E. (1913). *The marking system in theory and practice*. Baltimore, MD: Warwick & York Inc.

Flyvbjerg, B. (2004). Five Misunderstandings about case–study research. In C. Seale, G. Gobo, J. F. Gubrium, & D. Silverman (Eds.), *Qualitative research practice* (pp. 420–434). London: Sage Publications.

Flyvbjerg, B. (2006). Five misunderstandings about qualitative research. *Qualitative Inquiry, 12*(2), 219–245.

Frederiksen, J. R., & Collins, A. (1989). A system approach to educational testing. *Educational Researcher, 18*(9), 27–32.

Gee, J. P. (2003). Opportunity to learn: A language-based perspective on assessment. *Assessment in Education: Principles, Policy & Practice, 10*(1), 27–46.

Gipps, C. (1999). Socio-cultural aspects of assessment. *Review of Research in Education, 24*, 355–392.

Gipps, C. (2005). Commentary on "the idea of testing: Psychometric and sociocultural perspectives". *Measurement, 3*(2), 98–102.

Gipps, C., & Murphy, P. (1994). *A fair test? Assessment, achievement and equity* Buckingham, UK: Open University Press.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Chicago, IL: Aldine Publishing Company.

Gordon Commission on the Future of Assessment in Education. (2013). *To assess, to teach, to learn: A vision for the future of assessment. Technical report*. Princeton, NJ: ETS/Gordon Commission.

Gordon, M. E., & Fay, C. H. (2010). The effects of grading and teaching practices on students' perceptions of grading fairness. *College Teaching, 58*(3), 93–98.

Graham, P. (2005). Classroom-based assessment: Changing knowledge and practice through pre-service teacher education. *Teacher and Teacher Education, 21*, 607–621.

Green, S., Johnson, R. L., Kim, D. H., & Pope, N. S. (2007). Ethics in classroom assessment practices: Issues and attitudes. *Teaching and Teacher Education, 23*, 999–1011.

Guba, E. G., & Lincoln, Y. S. (1994). Competing paradigms in qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 105–117). Thousand Oaks, CA: Sage Publications, Inc.

Gummer, E. S. (2000). *Rhetoric and reality: Congruence between the knowledge and practice of assessment of the classroom science teacher and the reform initiative of the national science education standards.* (Ph.D. dissertation) USA: Purdue University. UMI No. AAT 3018203.

Guskey, T. R., & Jung, L. A. (2009). Grading and reporting in a standards-based environment: Implications for students with special needs. *Theory into Practice, 48*(1), 53–62.

Gynnild, V. (2011). Student appeals of grades: A comparative study of university policies and practices. *Assessment in Education: Principles, Policy & Practice, 18*(1), 41–57.

Heritage, M. (2013). Gathering evidence of student understanding. In J. McMillan (Ed.), *SAGE handbook of research on classroom assessment*. Thousand Oaks, CA: Sage Publications.

Heubert, J. P., & Hauser, R. M. (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: Board on Testing and Assessment, Commission on Behavioral and Social Sciences and Education, National Research Council.

Hollander, J. A. (2004). The social contexts of focus groups. *Journal of Contemporary Ethnography, 33*(5), 602–637.

Holtgraves, T. (2004). Social desirability and self-reports: Testing models of socially desirable responding. *Personality and Social Psychology Bulletin, 30*(2), 161–172.

Hughes, R., & Huby, M. (2004). The construction and interpretation of vignettes in social research. *Social Work & Social Sciences Review, 11*(1), 36–51.

Jackson, P. W., Boostrom, R. E., & Hansen, D. T. (1993). *The moral life of schools*. San Francisco: Jossey-Bass/John Wiley & Sons Inc.

James, M., & Pedder, D. (2006a). Beyond method: Assessment and learning practices and values. *Curriculum Journal, 17*(2), 109–138.

James, M., & Pedder, D. (2006b). Professional learning as a condition for assessment for learning. In J. Gardner (Ed.), *Assessment and learning* (pp. 27–43). London: Sage Publications.

Jensen, D. (2008). Transferability. In L. M. Given (Ed.), *The SAGE encyclopedia of qualitative research methods* (pp. 887–). Thousand Oaks: Sage Publications.

Jiménez, R. T. (2004/2005). More equitable literacy assessments for Latino students. In S. J. Barrentine & S. M. Stokes (Eds.), *Reading assessment: Principles and practices for elementary teachers* (pp. 49–51). Newark, DE: International Reading Association.

Joint Advisory Committee. (1993). *Principles for fair student assessment practices for education in Canada*. Edmonton, AB: Centre for Research in Applied Measurement and Evaluation, University of Alberta. Retrieved from http://www2.education.ualberta.ca/educ/psych/crame/files/eng_prin.pdf.

Joint Committee on Standards for Educational and Evaluation. (2003). *The student evaluation standards: How to improve evaluations of students*. Thousand Oaks, CA: Educational Policy Leadership Institute/Corwin Press Inc.

Kessels, J. P., & Korthagen, F. A. (1996). The relationship between theory and practice: Back to the classics. *Educational Researcher, 25*(3), 17–23.

Lane, S., & Silver, E. A. (1999). Fairness and equity in measuring student learning using a mathematics performance assessment: Results from the quasar project. In A. L. Nettles & M. T. Nettles (Eds.), *Measuring up: Challenges minorities face in educational assessment* (pp. 97–120). Norwell, MA: Kluwer Academic Publishers.

Laucius, J. (2009, April). Students aren't failing, but system is, teachers say. Educators feel pressured to ensure high schoolers pass; expert blames Ontario's drive to hike grad rates. *The Ottawa Citizen,* A1–A2.

Leibowitz, M. (2000). The work ethic and the habits of mind. In A. L. Costa & B. Kallick (Eds.), *Discovering and exploring habits of mind* (pp. 62–78). Alexandria, VA: Association for Supervision and Curriculum Development.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). *Complex, performance-based assessment: Expectations and validation criteria. CSE technical report 331*. Los Angeles, CA: Center for the Study of Evaluation.

Lovett, S., & Gilmore, A. (2003). Teachers' learning journeys: The quality learning circle as a model of professional development. *School Effectiveness and School Improvement, 14*(2), 189–211.

Martin, M., & Stella, A. (2007). *External quality assurance in higher education: Making choices* (Vol. 85). Paris: United Nations Educational, Scientific and Cultural Organization, International Institute for Educational Planning.

Maxcy, S. J. (2003). Pragmatic threads in mixed-method research in the social sciences: The search for multiple modes of inquiry and the end of the philosophy of formalism. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social & behavioral research* (pp. 51–89). Thousand Oaks, CA: Sage Publications.

Maxwell, J. A. (2005). *Qualitative research design: An interactive approach* (2nd ed.). Thousand Oaks: SAGE Publications.

McDonald, B., & Boud, D. (2003). The impact of self-assessment on achievement: The effects of self-assessment training on performance in external examinations. *Assessment in Education: Principles, Policy & Practice, 10*(2), 209–220.

McMillan, J. H. (2003). Understanding and improving teachers' classroom assessment decision making: Implications for theory and practice. *Educational Measurement: Issues and Practice, 22*(4), 34–43.

McMillan, J. H. (2011). *Classroom assessment: Principles and practice for effective standards-based instruction* (5th ed.). Boston, MA: Pearson Education Inc.

Merriam, S. B. (1998). *Qualitative research and case study applications*. San Francisco, CA: Jossey-Bass/John Wiley & Sons.

Messick, S. (2000). Consequences of test interpretation and use: The fusion of validity and values in psychological assessment. In R. D. Goffin & E. Helmes (Eds.), *Problems and solutions in human assessment* (pp. 3–20). Norwell, MA: Kluwer Academic Publishers.

Messick, S., & Anderson, S. (1970). Educational testing, individual development, and social responsibility. *The Counseling Psychologist, 2*(2), 80–88.

Monroe, W. S. (1923). *An introduction to the theory of educational measurements*. Cambridge, MA: Riverside Press/Houghton-Mifflin Company.

Moon, T. (2005). The role of assessment in differentiation. *Theory into Practice, 44*(3), 226–233.

Morgan, C., & Watson, A. (2002). The interpretive nature of teachers' assessment of students' mathematics. *Journal for Research in Mathematics Education, 33*(2), 78–110.

Moss, P. (2003). Reconceptualizing validity for classroom assessment. *Educational Measurement: Issues and Practice, 22*(4), 13–25.

Moss, P., Pullin, D., Gee, J. P., & Haertel, E. H. (2005). The idea of testing: Psychometric and sociocultural perspectives. *Measurement, 3*(2), 63–83.

Newkirk, T. (2004). A mania for rubrics. In A. S. Canestrari & B. A. Marlowe (Eds.), *Educational foundations: An anthology of critical readings* (pp. 199–201). Thousand Oaks, CA: Sage Publications.

OECD. (2011). *Lessons from PISA for the United States. Strong Performers and Successful Reformers in Education*. OECD Publishing. Retrieved from http://www.oecd.org/pisa/46623978.pdf.

Ontario College of Teachers. (2010). *Foundations of professional practice. Toronto, ON*. Retrieved from http://www.oct.ca/~/media/PDF/Foundations%20of%20Professional%20Practice/Foundation_e.ashx.

Ontario Ministry of Education. (2007). *The Ontario curriculum grades 9 and 10 English*. Toronto, ON: Queen's Printer for Ontario. Retrieved from http://www.edu.gov.on.ca/eng/curriculum/secondary/english910currb.pdf.

Ontario Ministry of Education. (2010). *Growing success: Assessment, evaluation and reporting in Ontario schools*. Toronto, ON: Queen's Printer for Ontario. Retrieved from http://www.edu.gov.on.ca/eng/policyfunding/success.html.

Parkes, J. (2007). Reliability as argument. *Educational Measurement: Issues and Practice, 26*(4), 2–10.

Parkes, J. (2013). Reliability in classroom assessment. In J. H. McMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 107–123). Thousand Oaks, CA: Sage Publications.

Patton, M. Q. (2002). *Qualitative research and evaluative methods*. Thousand Oaks, CA: Sage Publications.

Ploegh, K., Tillema, H., & Segers, M. (2009). In search of quality criteria in peer assessment practices. *Studies in Educational Evaluation, 35*, 102–109.

Poehner, M. E. (2011). Dynamic assessment: Fairness through the prism of mediation. *Assessment in Education: Principles, Policy & Practice, 18*(2), 99–112.

Popham, W. J. (2003). Seeking redemption for our psychometric sins. *Educational Measurement: Issues and Practice,* 45–48.

Popham, W. J. (2008). *Classroom assessment: What teachers need to know* (5th ed.). Boston, MA: Pearson Education Inc.

Pryor, J., & Crossouard, B. (2008). A socio-cultural theorisation of formative assessment. *Oxford Review of Education, 34*(1), 1–20.

Pryor, J., & Lubisi, C. (2002). Reconceptualising educational assessment in South Africa—Testing times for teachers. *International Journal of Educational Development, 22*(6), 673–686.

Resh, N. (2009). Justice in grades allocation: Teachers' perspective. *Social Psychology of Education: An International Journal, 12*(3), 315–325.

Rinsland, H. D. (1937). *Constructing tests and grading in elementary and high school subjects*. New York: Prentice-Hall Inc.

Robinson, J. M. (2002). In search of fairness: An application of multi-reviewer anonymous peer review in a large class. *Journal of Further & Higher Education, 26*(2), 183–192.

Russell, M. K., & Airasian, P. W. (2012). *CA: Concepts and applications* (7th ed.). New York: McGraw-Hill Higher Education.

Ryan, T. G. (2000). *An action research study of secondary science praxis.* (Ed.D. dissertation) Canada: University of Toronto. UMI No. AAT NQ58601.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119–144.

Schendel, E., & O'Neill, P. (1999). Exploring the theories and consequences of self-assessment through ethical inquiry. *Assessing Writing, 6*(2), 199–227.

Seidman, I. (1998). *Interviewing as qualitative research: A guide for researchers in education and the social sciences*. New York: Teachers College Press.

Shepard, L. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4–14.

Shepard, L. (2006). Classroom assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 623–646). Westport, CT: National Council on Measurement in Education, American Council on Education/Praeger Publishers.

Shulman, L. S. (1987/2004a). Knowledge and teaching: Foundations of the new reform. In S. M. Wilson (Ed.), *The wisdom of practice: Essays on teaching, learning, and learning to teach* (pp. 249–271). San Francisco, CA: Jossey-Bass.

Shulman, L. S. (1987/2004b). The wisdom of practice: Managing complexity in medicine and teaching. In S. M. Wilson (Ed.), *The wisdom of practice: Essays on teaching, learning, and learning to teach* (pp. 249–271). San Francisco, CA: Jossey-Bass.

Smith, E., & Gorard, S. (2006). Pupils' views on equity in schools. *Compare, 36*(1), 41–56.

Smith, J. K. (2003). Reconsidering reliability in classroom assessment and grading. *Educational Measurement: Issues and Practice, 22*(4), 26–33.

Speck, B. W. (1998). Unveiling some of the mystery of professional judgment in classroom assessment. In R. S. Anderson & B. W. Speck (Eds.), *Classroom assessment and the new learning paradigm* (pp. 89–96). San Francisco, CA: Jossey-Bass/John Wiley & Sons, Inc.

Stake, R. E. (2005). Qualitative case studies. In N. K. Denzin & Y. S. Lincoln (Eds.), *The Sage handbook of qualitative research* (3rd ed., pp. 443–466). Thousand Oaks, CA: Sage Publications.

Stiggins, R., & Chappuis, J. (2005). Using student-involved classroom assessment to close achievement gaps. *Theory into Practice, 44*(1), 11–18.

Stiggins, R., Frisbie, D. A., & Griswold, P. A. (1989). Inside high school grading practices: Building a research agenda. *Educational Measurement: Issues and Practice, 8*(2), 5–14.

Stiggins, R. J., & Bridgeford, N. J. (1985). The ecology of classroom assessment. *Journal of Educational Measurement, 22*(4), 271–286.

Stobart, G. (2006). The validity of formative assessment. In J. Gardner (Ed.), *Assessment and learning* (pp. 143–146). London: Sage Publications.

Stobart, G. (2005). Fairness in multicultural assessment systems. *Assessment in Education: Principles, Policy & Practice, 12*(3), 275–287.

Suskie, L. (2002). Fair assessment practices: Giving students equitable opportunities to demonstrate learning. *Adventures in Assessment14*. Retrieved from http://www.sabes.org/resources/publications/adventures/vol14/14suskie.htm.

Szpyrka, D. A. (2001). *Exploration of instruction, assessment, and equity in the middle school science classroom.* (Ph.D. dissertation) USA: University of Central Florida. UMI No. AAT 3029061.

Tata, J. (2005). The influence of national culture on the perceived fairness of grading procedures: A comparison of the United States and China. *Journal of Psychology, 139*(5), 401–412.

Tierney, R. D. (2008). *Fairness in classroom assessment: Multiple and conflicting interpretations*. Paper presented at the annual meeting of the Canadian Society for Studies in Education, Vancouver, 31 May.

Tierney, R. D. (2013). Fairness in classroom assessment. In J. H. McMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 125–144). Thousand Oaks, CA: Sage Publications.

Tierney, R.D., & Charland, J. (2007). *Stocks and prospects: Research on formative assessment in secondary classrooms*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Tierney, R. D., Simon, M., & Charland, J. (2011). Being fair: Teachers' interpretations of principles for standards-based grading. *The Educational Forum, 75*(3), 210–227.

Tierney, R. J. (1998/2005). Literacy assessment reform: Shifting beliefs, principled possibilities, and emerging practices. In S. J. Barrentine & S. M. Stokes (Eds.), *Reading assessment: Principles and practices for elementary teachers* (pp. 29–40). Newark, DE: International Reading Association.

Tomlinson, C. A. (2005). Grading and differentiation: Paradox or good practice? *Theory into Practice, 44*(3), 262–269.

Torrance, H., & Pryor, J. (1998). *Investigating formative assessment: Teaching, learning and assessment in the classroom Philadelphia.* PA: Open University Press.

Torrance, H., & Pryor, J. (2001). Developing formative assessment in the classroom: Using action research to explore and modify theory. *British Educational Research Journal, 27*(5), 615–631.

Volante, L. (2006). Reducing bias in classroom assessment and evaluation. *Orbit, 36*(2), 34–36.

Warms, C. A., & Schroeder, C. A. (1999/2009). Bridging the gulf between science and action: The "new fuzzies" of neopragmatism. In P. G. Reed & N. B. Crawford Shearer (Eds.), *Perspectives on nursing theory* (pp. 111–121). Philadelphia, PA: Lippincott Williams & Wilkins.

Warne, C. (2006). *Aristotle's Nicomachean ethics.* London: Continuum International Publishing Group.

Watson, A. (1999). Paradigmatic conflicts in informal mathematics assessment as sources of social inequity. *Educational Review, 51*(2), 105–115.

Weeden, P., Winter, J., & Broadfoot, P. (2002). *Assessment: What's in it for schools?* New York: RoutledgeFalmer.

Wendorf, C. A., & Alexander, S. (2005). The influence of individual- and class-level fairness-related perceptions on student satisfaction. *Contemporary Educational Psychology, 30*(2), 190–206.

Whittington, D. (1999). Making room for values and fairness: Teaching reliability and validity in the classroom context. *Educational Measurement: Issues and Practice, 18*(1), 14–22.

Wiggins, G. (1993). *Assessing student performance: Exploring the purpose and limits of testing.* San Francisco, CA: Jossey-Bass Inc.

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation, 37*(1), 3–14.

Wiliam, D. (1994). *Toward a philosophy for educational assessment*. Paper presented at the British Educational Research Association. Retrieved from http://www.kcl.ac.uk/depsta/education/publications/BERA _94.pdf

Williams, M. (2004). Generalization/generalizability in qualitative research. In M. S. Lewis-Beck, A. Bryman, & T. F. Liao (Eds.), *Encyclopedia of social science research methods* (pp. 421–422). Thousand Oaks, CA: Sage Publications.

Wilks, T. (2004). The use of vignettes in qualitative research into social work values. *Qualitative Social Work, 3*(1), 78–87.

Wilson, M. (2006). *Rethinking rubrics in writing assessment.* Portsmouth, NH: Heinemann Reed Elsevier Inc.

Wormeli, R. (2006). *Fair isn't always equal: Assessing and grading in the differentiated classroom.* Portland, MN: National Middle School Association.

Wyatt-Smith, C., Klenowski, V., & Gunn, S. (2010). The centrality of teachers' judgement practice in assessment: A study of standards in moderation. *Assessment in Education: Principles, Policy & Practice, 17*(1), 59–75.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing, 27*(2), 147–170.

Yin, R. K. (2006). Case study methods. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 3–31). Mahwah, NJ: Lawrence Erlbaum for the American Educational Research Association.

Yip, D. Y., & Cheung, D. (2005). Teachers' concerns on school-based assessment of practical work. *Journal of Biological Education, 39*(4), 156–162.

Yung, B. H. W. (2001). Three views of fairness in a school-based assessment scheme of practical work in biology. *International Journal of Science Education, 23*(10), 985–1005.

Zoeckler, L. G. (2005). *Moral dimensions of grading in high school English.* (Ph.D. dissertation) USA: University of Indiana. UMI No. AAT3183500.