English Teaching: Practice and Critique September, 2004, Volume 3, Number 2
http://education.waikato.ac.nz/research/files/etpc/2004v3n2art3.pdf pp. 48-73

Teachers working around large-scale assessment: Reconstructing professionalism and professional development

TREVOR GAMBELL
Department of Curriculum Studies, University of Saskatchewan, Canada

ABSTRACT: Educational reform initiatives are predicated on the professionalisation of teaching. Professionalisation implies that teachers assume and practice increased control in areas of non-instructional decision-making, rather than being preoccupied with content and procedural knowledge. As professionals, teachers are called upon to grapple with larger educational purposes and directions. In a more professional culture, teachers assume greater responsibility for generating their own expert knowledge. Large-scale assessment is often been portrayed as inimical to the interests of teachers and as an anathema by professional teacher associations. Little primary research has touched upon the impact of large-scale testing on teachers' self-identity, their sense of professionalism, and their use of evaluative research. This study examines why English teachers were motivated to take part professionally in a 1998 Canadian large-scale literacy assessment. Teachers were interviewed before, during, and four to six months after taking part in the scoring sessions. This paper examines an evolving and enhanced concept of professionalism among these teachers. Rather than robbing these teachers of their professional autonomy and judgment, participation in this assessment program challenged them to reflect upon their own content and practical knowledge, critique their own teaching and evaluation practices, and redefine their professional roles. Emergent themes include affirmation and reaffirmation, validation of knowledge and classroom practice, clarification of large-scale assessment's role in teacher and learning, enhanced professionalism, and philosophical shifts. These teachers reconstructed professionalism and undertook professional development on an individual and collective basis and through a variety of experiences considered inimical to the collective professional welfare of teachers just a few years ago.

KEYWORDS: Professionalisation, professionalism, (re)constructed professionalism, teacher knowledge, large-scale assessment, professional development.

INTRODUCTION

Large-scale assessment has been often portrayed in North America as inimical to the interests of teachers and as an anathema by professional bodies. For example, in a recent issue of the widely read *Phi Delta Kappan,* Alfie Kohn (2001) exhorted teachers to "make the fight against standardized tests our top priority…until we have chased this monster from our schools" (p. 349). Many teachers and their professional associations are predisposed to view large-scale testing programs with a jaundiced eye because they are often perceived as subverting teachers' professional right to evaluate their own students as an exclusive prerogative. State-mandated testing is frequently castigated in Canada too, as an external form of educational and social control that

disenfranchises the professional educator (Barlow & Robertson, 1994; Robertson & Ireland, 2000; Canadian Teachers Federation, 2000).

Large-scale "standardized testing", as teachers' associations abjure, is professionally debilitating in several ways because of its controlled design. First, critics argue, it devalues teachers' pedagogical skills in fostering learning in young people (Runté, 1998): "The greater the degree of curricular specificity dictated by the external examination, the more limited the teachers' need for, and claim to, professional autonomy." (p.167) Second, large-scale assessments are appendages of state or corporate authority (Robertson, 1998) and thus run contrary to the sense of individual empowerment and social mission that motivates many teachers in their practice. Third, the substantive and procedural aspects of testing substitute for, take precedence over, and frame that body of content and pedagogy which professionals assert is their right to control (Anderson et al., 1990; Wideen et al., 1997). And fourth, testing deskills teachers by usurping and supplanting their judgemental right of appraisal in evaluation.

However, as Cizek (2001) has pointed out, many claims about testing's deleterious impact are often unsubstantiated. In a wide ranging review, Mehrens (1998) found the evidence for purported negative consequences on either curricular content or instructional process to be skimpy; documentation supporting the conclusions about large-scale assessments' supposedly malignant influence on teacher motivation, morale, stress and ethical behaviour was likewise sketchy at best. Contrary, positive consequences have been described by Beaudry (2000) on Maine teachers' classroom assessment practices, by Goldberg and Roswell (1999/2000) on teachers' reflective and critical thinking when planning instruction, and by Bishop (1998, 2000) on North American student performance in an international context.

Within Canada, there has been little research into the impact of "low-stakes" assessments[1] among individual teachers, despite the introduction of such programs in the 1990s in many provinces and nationally. Nor has there been recognition of the extent to which teachers function as active agents in test construction, pilot testing, marking, and standards-setting in large-scale assessment (Lafleur & Ireland, 1999). In all provinces except one, it is teachers who develop and mark the tests. For example, in British Columbia, over 300 teachers develop test questions and approaches and over 1600 teachers are engaged in marking Grades 4, 7, 10 and 12 tests in any given year. Canada's only national assessment program, the School Achievement Indicators Program, was and is developed and marked by teachers (Council of Ministers of Education, Canada, 1999). Unlike the United States, where assessment is run largely by commercial organizations outside the education system, and Great Britain, where universities exercise the preponderate influence, ministries of education coordinate large-scale assessments in Canada. There, current assessment approaches demonstrate arms-length overview, linkages with classroom practice, and some concern with professional autonomy through process evaluation and anonymous sampling (Mawhinney, 1998).

[1] Low-stakes assessment is that which does not record or report individual student grades; all scores are aggregated. High-stakes assessment, on the other hand, records grades for individual students, and these grades form the evaluation for students in the subject area.

A fundamental issue, then, is whether such teacher involvement, as a variety of systematic inquiry (Cousins & Walker, 2000), promotes or erodes the professionalism of teachers in a climate of educational reform. For example, Locke (2001) asserts that teacher professionalism in New Zealand is being eroded by the heavy-handed educational reforms in that country, a phenomenon he calls "deprofessionalization". Doecke & Gill (2000-2001) write of the diminished concept of professionalism which has characterized reform movements in Australia, especially in the state of Victoria. Assessment in Canada comes in many forms with different purposes. However, a fundamental dichotomy can be drawn between "high-stakes tests", which are usually examinations directly yielding student grades reported at the individual, classroom or school level, and which affect student progress in school, in post-secondary education, or in their entry to the workplace. "Low-stakes assessments" are anonymous, random-sample exercises that protect the confidentiality of student, teacher and school. Because "low-stakes assessments" aggregate scores to provide provincial, national or international trends and profiles, there is no scrutiny and presumably less pressure on individual students and teachers. On the other hand, "low-stakes assessments" may actually be high-stakes for ministers of education, for governments, or for teachers' organizations who become publicly accountable for the results of student performance in schools.

The study reported here was animated by the following questions: How does participation as a scorer in a large-scale, low-stakes literacy assessment actually affect individual teachers in their sense of professionalism? Does active involvement as a scorer change their attitudes as autonomous professionals? How does participation affect teachers' skill in rendering evaluative judgements, and his/her instructional and assessment behaviour? What is the impact of direct involvement in scoring on teachers' professional identity and their sense of professionalism? This enquiry relies on the testimony of four teachers who scored the 1998 Canadian School Achievement Indicators Program (SAIP) literacy assessment materials, describes how it influenced their professional perspectives, and considers whether this participation has constituted professional development or professional debilitation (deprofessionalization).

FOUR ENGLISH TEACHERS

In a qualitative multiple case study, we[2] interviewed four teachers from a pool of approximately 80 Saskatchewan teachers who volunteered for and accepted a position to score the national reading materials for the Council of Ministers of Education Canada's (CMEC's) School Achievement Indicators Program (SAIP) testing in 1998. A total of 148 teachers, selected by the CMEC from each province, were involved. We chose these four informants by asking the national scoring leaders to randomly identify eight teachers: four females and four males, two each with urban and rural teaching experience, two each with recent (about five years) and longer (up to 20

[2] The study reported here is part of a larger, ongoing study of the voluntary involvement of teachers in large-scale assessments at provincial, national, and international levels in several subject areas, and their evolving concept of professionalization, teacher knowledge, and professional growth. The co-researcher is Darryl Hunter, recent Director of Student Assessment and Curriculum Evaluation, Ministry of Education, British Columbia, Canada.

years) teaching experience. We approached the first four who met these criteria and they agreed to participate in our study: one female with five years of urban experience, one female with 13 years of urban and rural experience, one male with six years of rural experience, and one male with 12 years of urban experience.

All four were English language arts middle years and secondary teachers. Participants were interviewed three times: prior to the scoring session in the spring of 1998 after recruitment; during the actual scoring session at the end of week one in the two-week July 1998 undertaking; and six months after the scoring session when teachers had recommenced their classroom teaching duties. (See Appendix for pre-scoring session, mid-session group, and post-scoring session interview questions). We conducted and tape-recorded all interviews. Interviews 1 and 3 were carried out individually. The second interview was a group event in an informal setting with all four informants, involving both researchers meeting after one of the days' scoring sessions in July. This group interview was also tape-recorded for later transcription. Our four teachers provided informed consent; pseudonyms are used throughout this paper.

During the two weeks of large-scale literacy assessment scoring sessions, all teacher-scorers were responsible for scoring approximately 25,000 thirteen-year-old (grade 8) and sixteen-year-old (grade 11) papers in reading collected nation-wide. Teacher scoring leaders first trained scorers in groups using scoring rubrics and exemplar papers, discussing and re-scoring until reliability within one point on a five-point scale was consistently reached. Then scoring of the papers began with regular checks for intra-rater and inter-rater reliability. If and when reliability deteriorated, re-calibration was undertaken either individually or as a group. As such, the marking exercise followed the principles common to most holistic scoring sessions found in North American large-scale assessments of literacy (White, 1985).

Ratna was a secondary English language arts teacher who had begun her career thirteen years previously in a small rural school. A year later, she began teaching in large, city high schools, finishing a postgraduate diploma in educational administration in 1993. Ratna is teaching half time as she raises her young family. She has been active with local and provincial teachers' associations, and has been involved in curriculum committees at both the local and provincial level. She has aspirations of finishing a master's degree and moving into educational administration. She had no previous experience with large-scale assessments and their administration, nor with scoring the materials. "I think change is just sort of the heart of my career. And I need change every couple of years to keep myself interested."

Ted had taught for a total of six years in the isolated north before moving to a city, when he was interviewed. Ted tries to keep up to date with research by reading professional journals and books. He incorporates such strategies as readers' and writers' workshop in his classroom. He has written for provincial professional journals, and has participated in a writing benchmarks project with his school board. He has also presented workshops at teacher conferences. Ted undertook with a colleague a provincial teachers' association-funded project involving action research on students' changing perceptions towards reading in a process approach classroom. He has also conducted in-services for grade 8 teachers in his school division to support implementation of a new English language arts curriculum. "I saw a lot of

changes in my students, a lot of positive changes in their attitudes."

Felicity was beginning her fifth year of teaching the middle grades when she was interviewed. She was beginning a master's degree in curriculum studies, and was on the English language arts implementation team in her school division, situated in a large urban area. Like Ted, she was involved in the writers' benchmark project with her city school division and found it a very positive professional experience. "My interest is evaluation and assessment and how to become better as a teacher. I don't think that there is anybody that is perfect at evaluating and any way that you can become better, I think, helps the students."

Kirk had been teaching for 12 years at both the elementary and high school levels, mainly in music, but his assignment had been exclusively English language arts in the previous two years. Kirk has a bachelor of music education degree, with teacher preparation also in English. He has made conference presentations, but has no previous experience with large-scale assessment or marking provincial examinations. For Kirk, participation in the national scoring session was "a good opportunity to learn perhaps new ways of scoring or marking or grading. I thought it would also give me a chance to see how my grading and my evaluation of students compares to what's being done in other places, perhaps."

BEFORE, DURING, AFTER: MICRO-EVOLUTION OF CHANGE

Findings are presented chronologically: the pre-scoring interview a month before the national marking exercise, the post-scoring interview approximately one semester after the SAIP scoring session in 1998, and the group interview midway through the scoring session. The interview questions focused on four aspects of professionalism which served as a framework for analyzing findings: personal and professional motivations for involvement in large-scale assessment; the perceived role of evaluation in classroom practice; perception of self as an autonomous professional; and the impact of evaluation on teachers' relationships with colleagues, students and administrators. For the purposes of this paper, findings are presented not as individual, fully developed and descriptive case studies, but rather as interpretive comparisons among the four informants. All interviews were transcribed *verbatim* and were read and approved by individuals.

Pre-scoring interviews

On the eve of the national scoring initiative, *Ratna* worried about the negative impact of large-scale assessment because it represented a judgement of teachers' competence; that is, teachers were being evaluated. Large-scale assessment was fear-provoking for teachers, and the way people react to fear is to reject something. "So I think typically that there is a negative attitude certainly toward assessment from what I've heard," she reported, "and I think the general negative attitude slips into the classroom teacher's." The provincial teachers' federation at the time was sceptical of national testing, and its stance coloured individual teachers' attitudes in the province toward assessment. Ratna saw her professional teacher organization as standing apart from and influencing classroom practitioners, rather than enabling classroom teachers to shape and drive the professional organization and its attitudes. (Ratna was quite

active locally in the teachers' association.) Ratna implied that the organization was not reflecting membership attitudes but was rather influencing them. She wanted to exercise her own professional judgment of the value of large-scale assessment by participating as a scorer.

She hoped that assessments were designed in such a way as to guide teachers "as individuals" by setting objectives and assisting them to meet those objectives. She believed this could be accomplished while setting and maintaining standards to move them forward nationally, and to address "the accountability issue". She recognized that standards and benchmarks were necessary for accountability, but she was concerned that teachers are vilified when results are interpreted as unsatisfactory achievement. Ratna deemed the SAIP as controversial. She needed to prove to herself that the SAIP couldn't be so one-sidedly negative. There had to be good reasons why large-scale literacy assessment was being done. Ratna hoped, as a teacher, that the national testing "was developed to give us some benchmarks and guidelines." She evinced an interest in measurement, wanted to take some evaluation tools away, and was excited about meeting teachers from across the country in a professional setting. Yet she wanted to evaluate the SAIP project on her own terms and experience it personally.

Ratna thus brought a judgmental purpose to the whole exercise; her primary motivation was to judge the scoring program. The main motive for her participation was moral resolution; is large-scale assessment good or is it bad? She wanted to know whether, "what you're hearing in the media or the way it's being interpreted has been wrong and you want to know [if] there's some [thing] positive going on here."

Ratna had an exceptionally coherent "philosophy of evaluation" in her classroom practice, assuming each student as unique with individual needs for which she sets individual goals. Classroom evaluation provides multiple opportunities for each student to demonstrate they are meeting those goals. At the same time, and perhaps in contradiction to her romantic belief that each student implicitly sets his or her own standard, she admitted doubts had arisen in her own ability to maintain consistency in holistic evaluation, which she also thought important in classroom practice. Rubrics and team scoring, she felt, would restore confidence in her evaluative procedures, a confidence that she lost while working in the isolation of the classroom setting. Ratna said, "I'm really hoping that I'm going to be inspired in terms of seeing what our students do know. I'm hoping that it's going to affirm that they are learning and that we are teaching. I think there's going to be some fulfillment in both ends." Rubrics are an appeal device, according to Ratna, for demonstrating fairness after her initial holistic marks have been assigned. Ratna used rubrics on those occasions "where you have to be objective and support your instinctively-assigned mark". Yet she was not primarily interested in the marking session as an exercise in collegial calibration, unlike some of her colleagues (for example, Kirk) who wanted to attune their marking practices with those of fellow professionals.

Ted, in contrast to Ratna, was dissatisfied with aspects of his classroom practice. He was knowledgeable about holistic scoring, but liked analytic scoring approaches because, in his terms, they offered more objectivity. Although Ted would call himself a whole language teacher who employs general impression scoring in language arts, at heart he is an analytic scorer. He asserted, "The more we can be objective instead of

subjective, the better we are." Ted's motives for participating in the SAIP initiative were: first of all, pecuniary; second, "fun" (which is a recurrent theme in his responses); and third, making connections with others who are like-minded. Ted was not generally interested in associating with teachers who are not "like-minded", whom he doesn't esteem as worthy professionals.

Whereas Ratna embodies some of the divisions and contradictions within the profession about testing, Ted highlights one professional viewpoint on or attitude toward large-scale assessment. Ted believes that large-scale assessments attract certain types of teachers like himself: professionals are those people who engage in curriculum implementation, large-scale assessment, and writing benchmarks programs. Interestingly, all these are officially sanctioned (by the ministry of education or the school division) initiatives, programs and undertakings. Ted's credo is that a professional subscribes to an initiative and becomes a teacher-leader through implementing it.

Another motive behind Ted's participation was professional development. Ted entered the national scoring session, not to clarify his own moral stance like Ratna, nor to become more efficient in marking like Kirk, but rather to develop comfort with holistic scoring in reading as a classroom application. Ted wanted to clarify his pedagogic values. He was not interested in scoring as a technique, but rather as an evaluative process that he could transport back to the classroom, or present at a workshop to other teachers. For Ted, large-scale assessment was an opportunity to reflect on one's own practice in student evaluation. Ted noted, "when we're looking at assessing students and looking at how they're doing, I think we're taking a look at ourselves and using it to make ourselves more effective." Large-scale scoring initiatives also model other processes, which can be transported back to the home community where he aspires to become a teacher-leader in implementation. Ted viewed himself as a leader through, and of, professional development. He saw assessment as a process tool for teachers to use; large-scale assessment scoring is a leadership training school in the application of that tool. Assessment for Ted was the educational issue of the times, and he foresaw the SAIP exercise as endowing him, perhaps positioning him, with leadership opportunities in demonstrating appropriate student evaluation practice. Ted also wanted to become a more effective evaluator, putting to use these tools and concepts in the classroom. But the word "effective" in his lexicon also meant being more enjoyable for students, since students learn through positive experiences. For Ted, the power of positive thinking and experience were the predominate traits of professional development.

Ted viewed himself as a mentor and a source of knowledge for colleagues. He was the popularizer of official ideas. Sometimes Ted equated being accountable with being more effective, which meant being more business-like, more professional-like, so that he can stand up and be counted among or above his peers. Ted was not looking for promotion but would accept it. Moreso, he was looking for approbation through people coming to seek him out for professional guidance. As he stated, "Many people with regards to literacy in my school come to me and ask, you know, for help with that."

Felicity's fundamental issue was her self-identified need to go outside her current and perceived narrow frame of professional reference for professional development. For

Felicity, it was a matter of broadening her horizons, of developing a broader frame of reference for her teaching and student evaluation. She stated, "With this type of exemplar and the rubrics, if I use them properly in my classroom, the kids are going to want to grow especially if you're using self-evaluation and peer evaluation. And I'm giving them immediate feedback which I don't [usually] do." Felicity made an implicit distinction between personal professional development and general professional development, the latter being scheduled, systematic, school division, or ministry of education in-service or curriculum implementation activities. Professional development, in other words, is that provided and organized by an agency. Felicity felt the need to go beyond immediate agencies for that development. Like Ted, she believed that participants in the scoring sessions would be like-minded teachers who exhibited "keen-ness".

These beliefs were central to her notion of an effective teacher: one who exhibits enthusiasm, high motivation, and a strong team spirit that means helping others and being helped in return. She did not believe that true professionals act autonomously. Yet she did not define professionalism in an organizational sense; professionals are individually motivated, but do not act autonomously. Like Ted, she recognized that there are teachers who are less than enthusiastic about their professional development, but felt she still could learn from the more experienced ones, whereas Ted was convinced he could not learn from the unlike-minded, regardless of their experience.

Felicity often used the word "focus" during this pre-scoring interview, when talking about evaluation and professional development. One focus was on learning outcomes where, curiously, she talked about evaluation as guiding and pre-scripting teaching. There was ambivalence, even contradiction, here. On the one hand, Felicity talked of evaluation as a means of charting a course in the vast sea of curriculum and instruction while, on the other hand, she found classroom evaluation to be an overwhelming burden. She did not feel comfortable with curriculum and instruction's role, seeing curriculum as the agent of the learning outcomes. These unresolved issues impelled her to participate in the national scoring session.

Kirk, in contrast to Felicity, foresaw the scoring venture as an opportunity for skills acquisition and calibration. Whereas Felicity was looking for expertise, Kirk was looking for external referents and techniques. He had been teaching English full-time for only two years. Although he was an experienced teacher, he felt like a newcomer when teaching English language arts. Yet Kirk was not looking for new evaluative processes, unlike Ted, but rather sought new forms of justification through the national project for his pre-existing student evaluation practices. He forecast the scoring session as functioning as a kind of benchmark or concurrent confirmation for classroom-assigned scores, whereas Felicity anticipated it as a type of social validation in the eyes of her teacher peers. Kirk thought that the relationship between instruction and evaluation to be arbitrary, vague, fuzzy, even haphazard. Clarification of the relationship between instruction and evaluation was his primary aim but, unlike Ratna, he was not concerned about adjusting objectives and teaching motivations using evaluative results.

For Kirk, the national exercise would be an extension of what he does in his classroom, whereas Felicity anticipated it would be an elaboration of what she has already learned through workshops conducted locally. Kirk saw professional

development as largely an individual experience, isolated and discontinuous, whereas Felicity was looking forward to the team activities. Kirk sought a larger frame of reference for his classroom marking. He wanted to meet other colleagues, and thereby adjust his marking with theirs to make it more "accurate", or more indicative of students' skills. Although he generally considered large-scale assessments as being politically instigated for public accountability, he didn't believe that schools were affected in any way by large-scale assessments (ostensibly because they were low-stakes). Kirk saw official documents as being the agent of change, not the act of participating in a professional development experience. In this way, he had a more bureaucratic view of professional development. Things are published and distributed, and it is up to the individual teacher to read and implement them. Like Ted, he was seeking expertise for eventual use in leadership.

Kirk was comfortable in the area of student evaluation, not discomforted. Kirk didn't really want his horizons expanded as Felicity did, but was more comfortable with his niche within the high school. He wanted to be among those who are aware, but not really involved. He wanted to position himself as a knowledgeable spectator on assessment issues, whereas Felicity wanted expertise. Kirk didn't want to become too deeply implicated in what could be a politically damaging initiative, if it were blacklisted by his teachers' federation, or not taken seriously by his colleagues. He foresaw this exercise as equipping him with knowledge that would demonstrate he was aware of current issues in evaluation.

In Kirk's terms, professional development meant career advancement, and the marking session would be a comprehension of evaluation techniques as well as an understanding of evaluation issues. But in the immediate term, Kirk was seeking congruence between grading and student ability. He wanted professional accuracy, in that he was looking for the true score. Kirk participated in the SAIP scoring because he wanted a means for justifying his marks to weaker students and to their parents. By devoting two weeks with colleagues from across the country to an examination of Canadian public school reading performance, he anticipated an external validation in measurement terms for his own classroom evaluation.

Post-scoring interviews

For *Ratna*, interviewed six months after the scoring session, the national marking session, in essence, meant a shift in "paradigms". She had set out before the experience with the aim of taking away some evaluation tools, and of re-establishing confidence in her own ability to maintain consistency in evaluation. Although she retains her discomfort with large-scale assessment, she reports that the scoring event was "amazing". She thinks that, "any time you have professionals talking like that, that's memorable and that's growth." She couldn't believe that so many teachers could adopt and consistently apply a uniform evaluation scheme to so many pieces of student work over such an extended period of time. It was beyond the scope of her imagination, because she believes so deeply in, personifies, and acts upon her notions of individuality and uniqueness. For Ratna, it is professionally reaffirming to have the scoring session move so many autonomous professionals to such "a common spot and common understanding", but she remains ambivalent about the impact of this exercise as professional development.

For her, large-scale scoring exercises make "a science out of an art". The act of training for uniformity and consistency entails such a shift in paradigm that it threatens mechanization of participants' thinking, which is anathema to Ratna. She had decided *a priori* that her fundamental values would not change, and neither did they; there are fundamental beliefs that she is "unwilling to abandon". At the same time, she acknowledges that when professionals get together, it provides an opportunity to affirm or re-affirm one's thinking. The resultant shifts of paradigm can create confusion, which can be a creative opportunity leading to professional growth. Ratna confesses that teaching affords "very few opportunities to affirm or re-affirm that what you're thinking and the confusion you're feeling is actually good." Professional development offers an opportunity to reconsider practice in a self-critical way. As she states, "You start to really reflect maybe more seriously on your views and beliefs." Comparing one's own practice and beliefs to those of others, especially younger teachers, offers a different mirror for reflection.

But at heart, seven months after the SAIP scoring, Ratna sees evaluation as an obligation (Ryan, 1997). She phrases this in terms of: "I'm a people person but once every couple of months I have to become a number cruncher." She resents and resists this; she does not want to become a more efficient, more accurate marker, as Kirk does. Both teachers see evaluation and learning as disconnected, but in very different ways, or for very different purposes. Ratna has anticipated diversity and is amazed to find such consistency in the teachers, male and female, from across the country. Ratna is sensitive to diversity, and was surprised at the commonality in approach achieved in scoring. Kirk, on the other hand, has approached the exercise presupposing a general commonality among students, but has come away from the marking sessions more attuned to the diversity of student experience and competence.

Has the experience widened Ratna's horizons? The question is, in her terms, "Could you teach an old dog new tricks?" The scoring experience has raised her awareness of the common issues faced by professionals across the country; yet participation hasn't changed her views about educational purpose and learning. In fact, she has consciously decided to try on a new conceptual paradigm, but just as consciously decided *a priori* that her underlying professional beliefs and values will not change or be corrupted by the testing initiative. Ratna had been looking for a professional transformation, to challenge her values, but this has not happened. Thus, ironically, there is a sense of disappointment that permeates her post-scoring commentary.

For *Ted*, professional development occurs where there is an opportunity to probe and develop pedagogical issues. He is concerned about student and parent perceptions of rubrics, while at the same time appreciates the rubrics as tools for professional reflection. He has returned home with readily adaptable tools and ideas for classroom practice and in-service. Ted thinks that the national scoring event has "had a positive impact because I've shared all the work I've done and some people have taken it and used it." Furthermore, the assessment experience has been a catalyst for further professional development. Whereas for many teachers, professional development is an episodic, discrete event in the context of a regulated school year, for Ted it is a continuous activity. He is self-directed and motivated, and was motivated at least in a minor way to participate in the scoring exercise for his own professional development. SAIP has served as catalyst for further professional activity, including writing about the session for a professional language arts journal.

According to Ted, the SAIP session has made a difference in how he approaches reading and its assessment. Initially, he had viewed rubrics and exemplars as having exclusively summative purposes, but now he can see the SAIP scoring procedures as having formative value when used with students. He has reconciled process approaches to teaching – he described himself as an holistic teacher – with assessment of the products of writing and reading produced by the students in his classroom(s). As well, participation permitted him to resolve the issues of subjectivity and objectivity in evaluation, which is akin to Ratna's distinction between the 'science and art' of teaching and evaluation. The SAIP scoring exercise has thus enabled Ted to reconcile a pedagogical issue relating to the compatibility of instruction and assessment. He reports that, "Now it's different in my classroom. After doing some research and some reading I realized, boy, it would be nice to make this more formative than summative and give the kids the opportunity."

Ted seeks out opportunities for professional development, and he sought out the scoring program specifically to enhance his professional stature in the eyes of colleagues in his own school division and beyond. Ted feels compelled to be a master teacher, and wants to be recognized as a forward-looking and exemplary educator. He looks for and lives on the affirmation of students, parents and other teachers. Yet Ted has not essentially changed in his view of student evaluation's importance. What may have changed are some classroom processes, and the passing on to students a greater role in self-evaluation and peer evaluation.

> I think students need to know where they're at and where they rank. And I think society demands that of us. And I've numerable arguments with people who would say different, but I believe society deems that we give grades and scores and I think it is necessary for us to do that. That's the way the world is now. And I think this is really important, you know, so I believe that you have to be fair and you have to have some criteria. And you have to have the whole shoe to match there.

Ted is no more knowledgeable about the impact of large-scale assessment on schooling and its macro structures at the end of the session than he was at the beginning. "I would hope that it effects it in a positive way. I hope that it changes curriculum and stresses the things we need to do Canada-wide. The primary impact is through research. It is reflection on practice. Research is a means for reflecting on practice."

For Ted, anything that changes classroom practice is a form of professional development. Yet he believes that professional development must be closely aligned with Department of Education curricular initiatives and assessment projects, or other reforms and initiatives by local school divisions. In that sense, Ted is actually somewhat conservative in how he aligns himself with established thinking.

The central, self-described metaphor for *Felicity's* involvement in the scoring session, and the focus for her desire for professional development, is her self-description as being "one little fish in the pond". Felicity is the least experienced of the four informants, and she participated to broaden her horizons. Felicity believes that professional development starts with the self – the individual has to want to change; the individual has to want to grow. Felicity admits to having learned a lot in the collaborative setting of the national scoring session. For her, professional development is not a solitary activity such as taking university courses. Felicity

believes that she "learned a lot professionally, working with the people that were in my group especially, actually." Professional development doesn't only mean what impacts on her, but also what impact she has on schooling, on others, and on teachers. Professional impact is that which happens in the classroom, but also involves what she can transact and share with others. Although she likes to learn from more experienced colleagues, she distances herself from those of the older guard whom she considers negative and cynical. As a newcomer to the profession, Felicity has recognized from the exercise that she does not have firm and fast beliefs about student evaluation's role, unlike the other informants.

For Felicity, professional development is primarily a social or collegial experience. She sees evaluative activity as primarily a communicative act, between teachers and students, teachers and parents, but especially among teachers. A scoring rubric has become a vehicle for better articulating expectations and communicating between teacher and parent. She has also come to recognize multiple types and purposes of assessment, whether in the classroom, large-scale assessment, program evaluation, or teacher supervision. She also sees the merit of "beginning with the end in mind", in terms of instructional planning. Felicity sees the potential for using assessment results to further curriculum implementation, to illuminate teachers' conferences, or to bolster educational leadership. She has come to see, like Ted, that students can play an active role in assessment, with rubrics and exemplars. Felicity asserts, "I think the exemplars and the rubrics make evaluation a lot clearer and a lot more concrete and we're getting away from it being so up in the air – it becomes more of a partnership then."

When interviewed six months after the session, Felicity has few substantive critical comments. She had received little preparation in student evaluation from her university pre-service training, so SAIP scoring provided her with an understanding of the subjectivity of scoring, about peer evaluation, and about being selective in one's marking. She describes her summer role in the marking initiative in terms akin to the student in the classroom. If SAIP scoring was a leadership training school for Ted, it was a two-week summer short course for Felicity. That was where she acquired the rudiments of evaluative techniques, within the supportive environment of peers, without the pressure of parents asking for the justification of a mark, and without the time constraints of day-to-day teaching. She has found that the large-scale assessment has enabled her to distinguish the important from the unimportant in expansive curriculum documents, but also to see the underlying rationale for some curricular approaches. Participation has deepened and focused her knowledge of curriculum in her province.

Kirk's motivation for participation in the summer scoring session was singular and definite: skills acquisition and calibration. He had wanted external justification for his existing evaluation practices, and he wanted to make evaluation more accurate, more efficient. And like Felicity, he too sought a larger frame of reference for his classroom practice, possibly because he had come to English teaching of late, after many years teaching music. But Kirk also wanted to be aware, to be up-to-date, and that requires being recognized as such by others including his peers.

As a professional development opportunity, the SAIP scoring session has affirmed Kirk's classroom practice. Kirk believes that SAIP scoring has enabled him to find

ways of "taking the subjectivity out of grading", which has value for public accountability purposes. He does believe that, for student work, exemplars and rubrics have primary value in enabling teachers and students to internalize the metric used for scoring. He thinks that assessment has a positive impact, particularly the scoring process, because it promotes objectivity, although group-scoring procedures might not be feasible in a high school. Hence, he views fairness in terms of consistency of standards for all students, and not as the alignment between assessment and objectives, as does Ratna. Kirk believes he is a better grader and teacher because the pieces of exemplary student writing enable him to tangibly demonstrate his marking criteria at home. Thus, he believes his scoring more accurately reflects actual student performance. He has posted the rubrics and exemplar papers on his classroom wall, as a first-order appeal device, and as a challenge to students in the quest for improved writing. Ratna, on the other hand, used exemplars as a second-order appeal device, and only in response to challenges to her grading.

Kirk has a very instrumental view of student evaluation, so it is not surprising that he has very instrumental views on his participation in the large-scale assessment initiative. He defines himself as a faster, more efficient grader now. He never lacked confidence as a classroom marker, unlike Felicity. He has not changed his underlying pragmatic value system, but rather has simply adopted the officially sanctioned criteria, as illustrated in the rubrics, for his classroom. Rather than having to undergo the longer and more complex process of devising his own evaluative criteria, he has rather opted to deploy those found in the large-scale assessment. So for Kirk, it is the increased efficiency in sorting that is important. Participation in the national scoring session has given him an accountability system at the classroom level, and has made him more aware of the full normative range of student performance. He notes, "I've learned a great deal about the kids at either end of the bell curve." But the scoring exercise has not fostered reflection on his own practice or values, or the practice of others. For Kirk, reflection means reinforcement for what he is already doing.

Kirk did profit from interactions with other teachers. The large-scale assessment scoring enabled him to commune with colleagues, but he does not think it affects curriculum or instruction. For Kirk, large-scale assessment remains a political enterprise to address the public demand for accountability; participation has had no impact in terms of his stature or status within the school setting. It has encouraged him to see evaluation as integral to student learning, but only in so far as the rubrics enable him to justify the marks he assigns and to enable students to self-evaluate. For Kirk, marking remains a professional technique that can be rendered more efficient, not a more value-laden exercise that involves introspection on his own professional values.

PROFESSIONAL DISCUSSION: GROUP INTERVIEW DURING THE SCORING SESSION

In this section, the findings are reported in general terms, as it was not consistently possible to identify individual voices in the tape recording of the interview.

Affirmation and re-affirmation

When interviewed during the national scoring session, participants affirmed how effectively and creatively the students had responded to the reading tasks. One respondent found it "enlightening" to see the variation in student responses. Another remarked on the exposure to a greater range of student ability than he had experienced at the classroom level, and that the provision of five levels of descriptors of student competence made it possible to discern a wider range of student reading ability. For the four interviewees, any doubts in students' ability to articulate their depth of reader response were dispelled by their marking experiences. The experience too had affirmed expectations about the assessment process, though one person noted that she would find it difficult to apply the process at the school level because of the time commitment required. One respondent noted that a "level of comfort" was reached when he was able to arrive at the sense that "what you're doing is literally for yourself as well as the project". This, it seems, occurs when a marker has internalized the process and is able to translate it into classroom use.

This understanding may arrive at different times throughout the scoring sessions and in different ways for different teachers. In fact, the conjuncture between individual and group professional purpose might never be reached for some teacher-scorers. In such cases, the national scoring session will not be experienced as professional development but merely as paid work. One respondent noted that she came to realize that the scoring process was "not that far from what I would have done anyway" and went on to state "I think that is a valuable thing". The speaker realized that the scoring process was both individually valuable, being validated through this national scoring exercise, but also that it was professionally of value to realize that one's classroom practice had been reaffirmed.

Another respondent likened the national scoring experience to graduate study, in that the experience led her to look "at things a lot differently now". This suggests that the experience provided the opportunity for reflecting on one's practice through comparing it with other models, such as the one presented at the national scoring session. This too seems a significant aspect of professional development, namely the opportunity to put one's practice up against other alternatives, models and possibilities.

Validation of classroom practice

A central issue for these teachers was to be able to justify the grade they assign student work in their classroom so as, in turn, to rationalize grades to students and to parents. The rubrics used for the scoring sessions, in which all scorers were trained and attuned for consistency, and against others for reliability, gave these teachers the confidence to assign grades that they considered valid, and were hence defensible, being derived through the scoring process with other experienced professionals. At least one respondent believed that grading is not compatible with learning, but that parents in particular are so driven by grades that finding a defensible and fair system of evaluation is needed. The hope for this teacher was that the national scoring session would provide an impartial system of categorizing student performances. In a sense, participants were able to feel professionally assured that the grade assigned was a professionally- and group-derived score, not merely an idiosyncratic one assigned in

isolation of other students' products, or of the expertise and experience of other educational professionals. Validation, as a social experience, meant a lot to these teachers, some of whom felt isolated in what they are doing at the classroom level. (This is a particular concern of rural teachers, many of whom are the only subject area teacher in their school.) One described it in these terms: "It really meant a lot of validity to what we were doing...where you're on an island by yourself and it's very frustrating sometimes." To consider one's self as part of a national community engenders a strong sense of professional satisfaction and validation.

One respondent became aware of gender bias in evaluation – she recognized that there was a subconscious tendency to mark the responses of females more generously than those of males, and to value more highly the attitudes of females as expressed in the tone and language choice in writing – through the national scoring exercise, that in turn had elicited thinking about the need for assessment tools to be carefully designed so as not to bias scoring along gender lines. It also prompted thinking about her own, perhaps subconscious, gender bias in evaluation.

Clarification of the role of large-scale assessment scoring in teaching and learning

An important realization for these four teachers was the difference between large-scale assessment scoring and student evaluation in the classroom. They recognized that this assessment was "low-stakes", and that student evaluation in their classroom is "high-stakes". As one said, "Not a single one of these marks is going to affect any students. It's designed to evaluate a school system, not students." They became aware of the issues around evaluation – educational issues, political issues. Implicit in their critique of "low-stakes" large-scale assessment was the belief that student performance is not a true indicator of ability because students have nothing to lose and thus do not perform optimally. For these teachers, student performance on such high stakes evaluations in the classroom was deemed a more accurate indicator of actual competence, because students were thought to be more highly motivated to perform well in a "high-stakes" situation.

These informants used the national scoring session to broaden their own evaluation skills. One noted that the experience provided a comparative basis for evaluation that would lead to broadening of his/her own classroom evaluation procedures. This respondent, for example, described how she found that lengthier responses weren't the most thoughtful and insightful ones. Being able to adapt the national scoring experience to classroom use was a tangible benefit for these teachers. One participant described how now he could become more detailed in the criteria used and feedback given at the classroom level, and how he would now "be able to break responses down a lot more". As well, respondents saw applications for student use, in that the rubrics provided could be made available for students to "recognize and almost grade themselves, and understand what's better, and how to get better and move from there." One teacher summed up this process of personalization and classroom application when she said, "It pretty much has to become a part of what you do when you leave."

None of these respondents counted external approbation, that is, recognition by school and district administrators, as the driving force for professional development. Rather,

they equated professional development with personal development; the rewards were primarily intrinsic and enhanced their own (personal) sense of professionalism.

Enhanced professionalism

Respondents noted that the benefits of participation in the national literacy scoring included meeting new people, making links with people in one's own community, sharing ideas, learning new things and developing professionally. Respondents did not perceive their involvement as professional coercion by others, including administrators, and at least one participant remained healthily sceptical (at this midway point in the exercise) about the process and being able to endorse it with his colleagues.

Another informant noted the important implementation opportunity that the experience provided, with upwards of 150 teachers returning to their home communities and being "able to influence innumerable students, perhaps over years and years." She, too, suggested that being able to take something away from the experience was a benefit of participation. At least two participants saw the potential for leadership by being among the first to transport, to their home school or division, an evaluative process that would potentially influence other professionals. This leadership role was not only an individual desire for administrative mobility, but also a way of enhancing professional status.

Apprenticeship was a sub-theme identified as related to professionalism. The less experienced teachers in this study had the sense of being novices among their more experienced peers, whom they viewed as seasoned and consummate professionals. One newer respondent expressed the realization she came to as a result of mingling with more experienced peers from across Canada.

> I myself must say, being a newer teacher, I found it more rewarding and it's built confidence in myself about evaluating, too, because you're checking how you feel on certain questions with someone that's taught for 25 years. Oh my gosh, they almost have exactly the same thing! You know, that kind of, that's real positive. [*Referring to her score compared with that of the more experienced teacher/scorers she was scoring with on the same student responses.*]

Negligible influence of large-scale assessment on literacy

These respondents remained sceptical about the impact of large-scale assessment results on literacy development in schools. They continued to see testing as a political initiative in the educational reform agenda, motivated for the wrong reasons by the wrong people (politicians). "It should affect it a great deal, but I don't think it does. And, I think that is the problem because it is a low-stakes thing.". This comment reinforces a previous point about the dubious generalizability of large-scale assessment data from students who know that the results have no bearing on their school grades. Another teacher defended the quality of current curricula suggesting that these curricula need to be properly and fully implemented before being considered for revision on the basis of large-scale assessment results.

This negative sense of the relationship between large-scale assessment findings and curriculum might be a function of their participation in a national assessment, where

curriculum development is a provincial/state responsibility. These teachers' opinions may change if and when they become involved in a provincial large-scale literacy assessment. Then the relationship between assessment results and curriculum becomes more transparent. Teachers generally will readily denounce grade 12 provincial exit examinations as driving curriculum, that is, teaching for the test. One wonders whether what grade 12 examinations influence is instruction rather than curriculum.

Philosophical uniforms

All educators practice from a philosophical heart, though in many cases they may not have reflected long enough on practice, or distanced themselves far enough, to contemplate what that philosophical heart is. Sometimes a professional development experience will provide the opportunity of time and/or distance that allows teachers to confront their own philosophical selves. When that occurs, they may face discomfort, not necessarily with their own professional selves, but with the philosophical premises behind the professional development experience. Then teachers are faced with a dilemma. Do they participate in the professional development experience and face the philosophical challenge, possible or implied, or coercion? Or do they turn aside from the experience and avoid possible discomfort and challenge?

One of the respondents in the study faced this dilemma during the exercise, and she speaks not only for herself, but for some other teachers who reportedly left the scoring session on the first day because, she believed, they were not prepared to work with the model of assessment presented. Here are her comments, framed in a clothing metaphor.

> If you want to go to work in a factory, and you choose to go and work in a factory you might have to wear a factory outfit. And if you don't want to work in a factory get another job. We all chose to be here and part of choosing to be here is we're going to accept their standards....you must accept the parameters and the philosophy by which this study is undertaken, and that is the premise. You have to accept that premise before you can go on here. You have to wear the outfit, whether it is professionally degrading or not.

Professional development opportunities such as large-scale assessment allow teachers to try on different philosophical uniforms, as well as look in the mirror themselves and examine their own pedagogical values.

CONCLUSIONS

Shulha & Cousins (1997) have identified three primary uses for evaluative information: instrumentally, to effect changes in program or practice; conceptually, to alter outlooks or ways of defining issues or problems for solution; and symbolically, to reify or undermine a pre-existing policy or position. In our study, we discerned a fourth use: as a valuation process to help practitioners clarify what is important and not important in their praxis. For those teachers we interviewed, participation in the 1998 national literacy reading scoring session enabled them, in fact, to use the evaluative information in all four ways. For Kirk, the evaluative processes provided instruments that altered and refined his classroom practice. For Ratna, the exercise

offered an opportunity to try operating within another conceptual paradigm. For Ted, SAIP scoring endowed him with a set of experiences that he could use to reinforce the symbolism of professional leadership. For Felicity, the venture more fundamentally shaped and clarified her professional values in student evaluation.

For all four participants, professional development (PD) was largely a communal process, occurring in a social and professional context, dealing with professional matters. Teachers often feel they are working in isolation (Ratna), that they need communally-established, hence validated, means of evaluation (Kirk), that they need to be on the cutting edge of evaluative (or curricular and instructional) processes (Ted), or that they seek validation through consort with like-minded professionals (Felicity). For all these teachers, moreover, participation affirmed or improved their classroom assessment practices, a shortcoming found repeatedly in pre-service education across North America (Daniel & King, 1998; Impara & Plake, 1996; O'Sullivan & Chalnick, 1991; Stiggins, 1999). In short, a variety of learning opportunities for teachers are embedded within the exercise (Falk & Ort, 1998).

Did assessment construct curriculum and teacher practice?

None of the informants accused the scoring exercise as detracting from their teaching. In all cases, participation in the "low-stakes", large-scale assessment reinforced teachers' classroom work rather than undermined it. By imparting a sense of confidence, by enhancing leadership training, by validating classroom practice, by supplying an external referent for a professional's classroom judgements, and by providing a collegial forum wherein teachers could reflect on their classroom practice, this scoring session supported teacher development. According to these four teachers, this low-stakes CMEC assessment project was not deemed to be professionally corrosive, although they professed doubts about the potentially negligible or negative impact after scores were released, especially if results were released in ways that compared and contrasted jurisdictions.

The professional issues encountered by teachers were not earth-shaking in their portent, nor had the national scoring event the character of an epiphany in their teaching lives. Rather, the exercise was another milestone in their professional development. As a crucible wherein values are clarified, a large-scale scoring session appears to affirm rather than undermine teachers' sense of professionalism. The four participants here generally saw the marking session as consistent with, rather than throttling, their instructional and assessment behaviour.

I suggest that the assessment tasks, the scoring rubrics, and the exemplar papers did not construct these teachers either theoretically or in terms of practice except in so far as each participant sought out the scoring experience for what it could possibly offer them, either philosophically or instructionally. Ratna is a good example of a philosophical construct that she wanted tested against the scoring process because she saw conflicts in her own philosophy of teaching where each student is the benchmark for his or her own development and learning. Ratna wondered how large-scale massed assessment could possibly offer practices compatible with her philosophy of individualism. For Ratna, the result was a shift in paradigms, which although entailing confusion, ambiguity and discomfort, also provided creative opportunity for practice that enabled her to retain her basic philosophy while yet embracing diverse

teaching practices. Ratna was looking for and experienced professional transformation wherein her own views were challenged.

Ted considered professional development (PD) the chance to probe and develop pedagogical values; his interest was an instructional one. With Ted, PD is anything that changes classroom practice, as long as it fits with official curriculum initiatives and assessment processes. Ted's views of curriculum were prior constructed and invariable, thus the scoring experience did nothing to subvert his beliefs (and neither did it subvert Ratna's philosophy). The holistic scoring regimen fitted Ted's espoused holistic philosophy of English language arts teaching, so there was no paradigm shift. Unlike Ratna, Ted is unlikely to seek out PD opportunities that do conflict with his philosophy.

Felicity wanted to broaden her frame of reference for English language arts teaching, and to do so by consorting with her more experienced peers. Felicity believes that PD begins with the self, who has to have the desire to change. Unlike Kirk, PD for Felicity is a social and collegial experience. Kirk, on the other hand, considers PD as largely an individual and episodic experience, isolated and discontinuous. It offered an opportunity for skills acquisition; issues of curriculum were non-existent, and instruction was affected secondarily and functionally through applied skills of assessment. Kirk looked for external justification for his existing practices.

Quite clearly, large-scale scoring events are episodic in teachers' professional lives. "Low-stakes", large-scale assessments in their administration and reporting remain peripheral at best to teachers' classroom concerns. Participants did not see the School Achievement Indicators Program as affecting curriculum except in neutral or positive ways. What professionals do with and within an evaluation may be as important as what an external evaluation "does" with professionals, though all four teachers in this study had their own reason for taking part. I would conclude that the scoring experience, being low-stakes and voluntary, had little impact on their philosophical attitudes toward curriculum, and practical rather than theoretical impact on praxis. Teacher scoring within large-scale assessments may not be catalytic or cataclysmic in its effects, but neither is it anathema to teachers' professional development.

Emergent issues

A number of issues emerge from these interviews that help to characterize the nature of teacher professionalism and of reconstructed professional and professional development. I think it is clear that teachers define professionalism in their own ways, and that these ways steer them into professional development experiences through a variety of self-determined reasons and purposes. The first issue arising from this study is that of PD as either an individual experience (Kirk) or a social experience and group (Felicity). Possibly this orientation varies with teacher experience; newer teachers may well opt for group PD experiences that offer the opportunity for professional cohesion, apprenticeship and mentoring rather than validation of practice or questioning of philosophy. A second issue is that of challenge and affirmation. Does one enter a PD event with the rather daunting and risky intent, like Ratna, of putting one's own experience up against other alternatives, models and possibilities, or is motivation driven by affirmation of practice and belief, as for Ted? Intent is not necessarily tied to years of experience; Ted and Felicity were

both relatively inexperienced teachers.

A third issue is that of motivation for participation in PD opportunities, characterized by intrinsic and extrinsic factors. Ratna and Felicity were impelled by their desire for personal growth and personal development; hence, there was no expectation of administrative approbation or for peer recognition. Ted and Kirk were in search of tools for career advancement, and peer recognition in terms of leadership roles and knowledge sharing were important external factors for both of them. (Despite Kirk's asserting that neither his school principal, nor most of his peers, were aware or necessarily interested in his participation in the scoring session, Kirk's career advancement sights were on a higher plane, namely the school division level.) Interestingly, none of the four voiced their belief that external approbation or recognition by school or district administrators was important, but peer recognition was prized. It was considered a badge of honour to have been selected to participate in the scoring session, and to have had their names put forward by their teachers' association.

Linked with the above issue is a fourth one, namely enhanced professionalism and professional status among peers. Although enhanced professionalism is a goal of all four participants, it was less so for Ted and Kirk than for Ratna and Felicity. This is an interesting issue, because it exemplifies the individual nature of constructed professionalism. To identify oneself as a professional implies a sense of confidence in one's abilities to practice the profession of teaching, and such confidence seems to be sustained intrinsically for some teachers, but differentially by extrinsic relationships for others. A fifth issue is that of apprenticeship and leadership. It seems that for some participants, the acquisition of skills and knowledge enables a shift from being an apprentice to becoming a leader. This is a particularly interesting issue, because it suggests that those who pursue PD to gain knowledge or attain skills for external purposes may well embrace a knowledge-based, hierarchical approach to leadership – a leader needs to be the most knowledgeable member of the school (or educational unit) and leads others through dissemination of knowledge. This seems to be in contrast to a facilitative philosophy of leadership, where an administrator will recognize and encourage individual members who have strong knowledge and skills bases, and facilitate leadership diversely among peers. The first approach centres the leader among her or his peers but in an hierarchial relationship; the second approach creates a peripheral role for the facilitator and positions all involved in a more equal relationship.

In terms of professionalism, two points became clear to me that seem to cut across all four teachers in this study. Ratna, Ted, Felicity and Kirk came to the scoring experience with a constructive scepticism and a critical perspective, to various degrees. These pre-requisite positionings, I believe, are an important characteristic for the enaction of reconstructed professionalism, and was vividly put into words in the factory uniform metaphor. The second point is the unexpected (by me) lack of a relationship between the scoring experience and curriculum. For me it begs the question: if these teachers entered with a critically analytic optic, why did they not construct their own implications for curriculum? I believe that the answer lies in the fortuitous timing of the assessment project. New English language arts curricula had been recently put in place in Saskatchewan, and teachers were involved in the curriculum development, piloting and implementation process along the way. The

testing and scoring processes were philosophically in empathy with the new curricula. Possibly most important was that this assessment was low-stakes for students, and hence for teachers. Unlike their international peers in the USA and UK, these teachers were not forcefully confronted with philosophical and paradigmatic cataclysms that required compliance or rebellion. Professionalism was not put to the test. Possibly these teachers were motivated to take part in the assessment scoring because they were triggered by the recent new curricula and sought opportunities to engage with practices that promised to mesh philosophically with the new curricula. If so, such self-determined professionalism is evidence of reconstruction.

REFERENCES

Anderson, J., Muir, W., Bateson, D., Blackmore, D., & Rogers, W. (1990). *The impact of provincial examinations on education in British Columbia: General report.* (Report to the British Columbia Ministry of Education.) Victoria, BC: British Columbia Ministry of Education.

Barlow, M., & Robertson, H-J. (1994). *Class warfare: The assault on Canada's schools.* Toronto: Key Porter Books.

Beaudry, J. (2000, April). *The positive effects of administrators and teachers on classroom assessment practices and student achievement.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Bishop, J. (1998). The effect of curriculum-based external exit exam systems on student achievement. *Journal of Economic Education, 29*(2), 171-182.

Bishop, J. (2000). Curriculum-based external exit exams: Do students learn more? How? *Psychology, Public Policy, and Law, 6*(1), 1999-215.

Canadian Teachers' Federation. (2000). *Standardized tests + High stakes = Educational inequity.* Retrieved June 17, 2002 from http://www.ctf.ca/e/what/other/assessment/high_stakes.htm

Cizek, G. (2001). More unintended consequences of high stakes testing. *Educational Measurement: Issues and Practice, 20*(4), 19-27.

Council of Ministers of Education, Canada (1999). *School Achievement Indicators Program: 1998 Reading and Writing Assessment, 13- and 16-year-old students.* Toronto: Council of Ministers of Education, Canada. Retrieved September 29, 2004 from http://www.cmec.ca/saip/rw98le/pages/tablee.stm

Cousins, J., & Walker, C. (2000). Predictors of educators' valuing of systematic inquiry in schools. *Canadian Journal of Program Evaluation,* Special Issue, 25-52.

Daniel, L., & King, D. (1998). Knowledge and use of testing and measurement literacy of elementary and secondary teachers. *Journal of Educational Research, 91*(6), 331-344.

Doecke, B., & Gill, M. (2000-2001). Setting standards: Confronting paradox. *English in Australia, 129 & 130,* 5-16.

Falk, B., & Ort, S. (1998). Sitting down to score: Teacher learning through assessment. *Phi Delta Kappan, 80*(1), 59- 64.

Goldberg, G., & Roswell, B. (1999/2000). From perception to practice: The impact of teachers' scoring experience on performance-based instruction and classroom assessment. *Educational Assessment, 6*(4), 257-290.

Impara, J., & Plake, B. (1996). Professional development in student assessment for educational administrators: an instructional framework. *Educational Measurement: Issues and Practice, 15*(2), 14-19.

Kohn, A. (2001). Fighting the tests: A practical guide to rescuing our schools. *Phi Delta Kappan, 82*(5). 349-357.

Lafleur, C., & Ireland, D. (1999). *Canadian and provincial approaches to learning assessments and educational performance indicators*. Ottawa: Canadian International Development Agency.

Locke, T. (2001). Questions of professionalism: Erosion and reclamation. *CHANGE: Transformations in Education, 4*(2), 30-50.

Mawhinney, H. (1998). Patterns of social control in assessment practices in Canadian frameworks for accountability in education. *Educational Policy, 12*(142), 98-109.

Mehrens, W. (1998). Consequences of assessment: What is the evidence? *Educational Policy Analysis Archives*, 6(13). Retrieved 30 September, 2004 from http://epaa.asu.edu/epaa/v6n13.html

O'Sullivan, R., & Chalnick, M. (1991). Measurement-related course requirements for teacher certification and recertification. *Educational Measurement: Issues and Practice, 10*(1), 17-19, 23.

Robertson, H-J. (1998). *No more teachers, no more books: The commercialization of Canada's schools*. Toronto, ON: McClelland & Stewart.

Robertson, H-J., & Ireland, D. (2000). *Form and substance: Critiquing SAIP*. Paper presented at the Annual Conference of the Canadian Society for the Study of Education, University of Alberta, Edmonton, AB, Canada, May 25, 2000.

Runté, R. (1998). The impact of centralized examinations on teacher professionalism. *Canadian Journal of Education 23*(2), 166-191.

Ryan, A. (1997). Professional obligement: A dimension of how teachers evaluate their students. *Journal of Curriculum and Supervision, 12*(2), 118-134.

Shulha, L., & Cousins, J. (1997). Evaluation use: Theory, research, and practice since 1986. *Evaluation Practice, 18*(3), 195-208.

Stiggins, R. (1999). Evaluating classroom assessment training in teacher education programs. *Educational Measurement: Issues & Practice, 18*(1), 23-27.

White, E. (1985). *Teaching and assessing writing*. San Francisco, CA: Jossey-Bass.

Wideen, M., O'Shea, T., Pye, I., & Ivany, G., (1997). High stakes testing and the teaching of science. *Canadian Journal of Education 22*(4), 428-444.

APPENDIX

Pre-scoring session interview questions

1. Please describe your career in education thus far, including coeducational experiences.
2. What are your major teaching subject areas? What was your academic preparation?
3. What are your previous experiences with large-scale assessments and provincial examinations?
4. Why did you apply for a position as national assessment scorer?
    a. What made you decide to complete the application form?
    b. Are you concerned about having the needed experience or skills?
    c. What is your understanding of national and provincial assessments?
    d. What is your understanding of what you will do during the national scoring sessions?
5. What relationship do you see between large-scale assessments, either provincial or national, and your classroom teaching?
    a. Do you think the assessments have a positive, neutral or negative impact on classroom teaching and learning? How so?
    b. What is your understanding of how (the processes by which) national and provincial assessments are developed?
    c. What is your understanding of why provincial and national assessments are created?
6. How do you think that large-scale scoring and classroom marking are related?
    a. What is your understanding of how large-scale assessments are marked?
    b. What effect do you think the summer experience will have in how you will mark your students next fall?
    c. What effect do you think the summer experience will have in how you relate with your teaching colleagues? With school administrators? With the parents of the students you teach?
7. What do you think about the role of evaluation and teaching?
8. What do you think about the role of evaluation and student learning?
9. What do you think will be the benefits of participating in the scoring session? What do you think are some potential drawbacks to participating?
10. What effect do you think that the summer scoring experience will have on your teaching? On your school? On your fellow teaching colleagues?
11. In general, how do you think large-scale assessment affects curriculum and instruction?
12. In general, how do you think large-scale assessment affects classroom evaluation?
13. In general, how do you think large-scale assessment affects teachers as professionals?
    a. What effect will participation in the summer scoring sessions have on your community status as a teacher?
    b. What effect will participation in the summer scoring sessions have on your stature with your teaching colleagues? How will your colleagues see you?

c. What effect will participation in the summer scoring sessions have on your stature with your employers or superiors? How will your principal, director, board perceive you?

14. How does large-scale assessment affect literacy and its teaching?
15. How would you describe your current knowledge about student evaluation?

Group session interview questions

1. How are you finding the national scoring exercise here in Saskatoon?
2. What have you learned so far about assessment?
3. What have you learned from your colleagues across Canada?
4. How would you describe your current knowledge about student evaluation?
5. What relationship do you see between large-scale assessments and your classroom teaching?
    a. Do you think the assessments have a positive, neutral or negative impact on classroom teaching and learning? How so?
    b. What is your understanding of how (the processes by which) national and provincial assessments are developed?
    c. What is your understanding of why provincial and national assessments are created? How are they used?
6. How do you think that large-scale scoring and classroom marking are related?
    a. What have you learned about how large-scale assessments are marked?
    b. What effect do you think the summer experience will have in how you mark your students next fall?
    c. What effect do you think the summer experience will have in how you relate with your teaching colleagues? With school administrators? With the parents of the students you teach?
7. What do you think about the role of evaluation and student learning?
8. What do you think will be the benefits of participating in the scoring session? What do you think are some potential drawbacks to participating?
9. What effect do you think the summer scoring experience will have on your teaching? On your school? On your fellow teaching colleagues?
10. In general, how do you think large-scale assessment affects curriculum and instruction?
11. In general, how do you think large-scale assessment affects teachers as professionals?
    a. What effect will participation in the summer scoring sessions have on your community status as a teacher?
    b. What effect will participation in the summer scoring sessions have on your stature with your teaching colleagues? How will your colleagues see you?
    c. What effect will participation in the summer scoring sessions have on your stature with your employers or superiors? How will your principal, director, board see you?
12. How does large-scale assessment affect literacy (English language arts) and its teaching?

Post-scoring session interview questions

1. Looking back eight months, what are your reflections on the national scoring exercise in Saskatoon? Would you do this again?
2. What have you learned about assessment?
3. Do you think the assessments have a positive, neutral or negative impact on classroom teaching and learning? How so?
4. What relationship do you now see between large-scale assessments and your own classroom teaching?
5. What is your current knowledge about student evaluation?
6. How do you think that large-scale scoring and classroom marking are related?
7. What effect did last summer's experience have in how you now mark your students?
8. What impact did the national assessment experience have in terms of your relationships with your teaching colleagues? With school administrators? With the parents of the students you teach?
9. What effect did participation in the summer scoring session have on your community status as a teacher?
10. What do you now think about the role of evaluation and student learning?
11. What do you now think were the benefits of your participating in the scoring session? What do you think were drawbacks to your participating?
12. In general, how do you think large-scale assessment affects curriculum and instruction?
13. In general, how do you think large-scale assessment affects teachers as professionals?
14. How does large-scale assessment affect literacy (English language arts) and its teaching?

Supplementary themes for exploration

15. Who is responsible for professional development: individual initiative, professional organization, board employer or distant agency such as Saskatchewan Education?
16. Does participation in a large-scale scoring session differentiate/distance you form other teachers, or does it enable you to better work, and build networks with other teachers?
17. Has the experience created a professional development focus for you? If so, what is that? Retrospectively, does that focus match your original intent for becoming involved?