Professionalism and High-Stakes Tests: Teachers' Perspectives When Dealing With Educational Change Introduced Through Provincial Exams

Carolyn E. Turner

The effect of high-stakes tests on classroom activity (commonly called washback) is an issue that is receiving heightened attention in the literature. It is yet one more element that teachers need to deal with in their professional contexts. This article focuses on the perspectives of ESL secondary teachers as they experience curriculum innovations introduced into the educational system via provincial exams. Survey results from 153 teachers are reported. The survey is part of a larger washback study that also triangulated classroom observation and teachers' and students' perception data in a longitudinal study. The survey results suggest that teachers would like to do their part in moving the system into a position where curriculum, their teaching and assessment, and the system's high-stakes exam correspond. They achieve this, however, according to their beliefs and professional stances, which may not present a unified performance across teachers.

Les chercheurs se penchent davantage sur l'effet qu'ont les examens à enjeux élevés sur l'activité en salle de classe (connu sous le nom de saut arrière). Il s'agit d'encore un autre élément dont doivent tenir compte les enseignants dans le cadre de leur travail. Cet article porte sur les perspectives d'enseignants d'ALS au secondaire qui sont confrontés à des innovations aux programmes scolaires introduites par les examens du ministère. Une enquête a été entreprise auprès de 153 enseignants dans le contexte d'une étude longitudinale portant sur le saut arrière où l'on a triangulé des données découlant d'observations en salle de classe, d'une part et des perceptions des enseignants et des élèves, d'autre part. Les résultats donnent à penser que les enseignants aimeraient contribuer à faire évoluer le système de sorte à faire correspondre leur enseignement et leur évaluation, les programmes scolaires, et les examens à enjeux élevés. Toutefois, la contribution des enseignants reposerait sur leurs croyances et leurs attitudes professionnelles, ce qui pourrait ne pas être uniforme d'un enseignant à l'autre.

Introduction

Second-language teachers have much to contend with serving as professionals in an ever-changing context of student populations, curriculum, and

classroom practice. One further element to deal with, which is receiving heightened awareness in the recent literature, is the impact of high-stakes tests on classroom activity.

This article focuses on ESL secondary teachers as they experience curriculum innovations introduced into the educational system via provincial exams. It reports on the perspectives of teachers as professionals in this situation. The results reported here are part of a larger longitudinal study entitled *Investigating high-stakes test impact at the classroom level*. The general research question for the larger study is: Does the involvement of the Ministry of Education, teachers, and students at various stages of the testing cycle make a difference in promoting beneficial washback in terms of teaching methodology and content, classroom testing methodology and content, participant perceptions, and student learning strategies? OR, is negative impact observed? Analysis of the data is ongoing. To date only the initial research design and preliminary findings have been presented (Turner, 2002, 2005). The sources of data were classroom observations, participant interviews, teacher discussions, case-study questionnaires, and a program-wide teacher survey.

This article specifically focuses on a program-wide teacher survey and deals with two concepts: washback and professionalism. The characteristics of their relationship emerged from the data and are reported here through an analysis of teacher questionnaire results. Before going any further, explanations and definitions are in order.

The phenomenon of the influence of tests on classroom activity is commonly referred to as *washback*. In educational systems, washback can affect students, teachers, parents, and ministries of education and other stakeholders.

One form of washback is related to innovation theory (Wall, 2000). Various actions and consequences may occur when an educational system wants to make changes (innovations) to a program. There are many ways to go about this. For example, a new official curriculum or program can be developed and presented. Another way (which may happen while waiting for a new curriculum to become official) can be to introduce the new procedures or content into the system through high-stakes exams. This is done in the hope that teachers will change or align their instructional practices to correspond to the exam materials and methodology. Teacher information sessions are sometimes offered to help with this process. Henrichsen (1989) discusses employing high-stakes tests in this manner as one way to enhance reform in a system. Drawing on general and language education literature, Andrews (2004) discusses in detail the relationship between washback and curriculum innovation.

In this article, the specific definition of washback is the extent to which the test influences language teachers and students to do things they would not

necessarily otherwise do (Alderson & Wall, 1993). In other words, the effects are only washback evidence if they can be linked to the introduction and use of the test (Messick, 1996). The terms *washback* and *test impact* are used interchangeably, although some places in the literature make a clear distinction between washback on local effects and test impact on societal effects (McNamara, 1998)

A precise definition of the second term, *professionalism*, remains elusive in the literature. As stated in the recent special issue of *TESL Canada Journal* (2004), it seems to be a complex construct with little academic literature (Mathews & Chuntian, 2004, p. i). If one looks further, however, definitions do appear that are specific to a context or study. In combination, they begin to provide a clear picture. Mathews and Chuntian use the *Canadian Oxford Dictionary* (1998) definition, "the skill or quality required or expected of members of a profession … one that involves some branch of learning or science." Englund (1996) includes the importance of requisite traits and functions (e.g., teacher training, ongoing professional development), but emphasizes the internalization of these events as individual teacher characteristics of professionalism: "The internal quality of teaching" (p. 77). Hedgcock (2002) expands on this and focuses on the reflective nature of teachers, viewing professionalism as the ability to think critically about practice, rather than relying on mechanical teaching strategies and methods. Kumaravadivelu (2003) sees teacher programs as being responsible for creating a climate of professionalism and for helping to develop teachers to "acquire the necessary knowledge, skill, authority, and autonomy to construct personal pedagogic knowledge" (p. 42). In the literature, professionalism does not appear to take on a unified definition; instead, one is able to weave together a multifaceted concept from specific instances. At the end of this article, the concept is revisited and expanded in the light of what characteristics emerged from the data in relation to teacher perspectives on dealing with educational change introduced through provincial exams.

I first provide background on the concept of washback in general and in second-language (L2) education in particular. This will include the necessity to hear the voices of teachers, who are main stakeholders, concerning their students' performance on externally developed high-stakes exams. I next mention a longitudinal study in the province of Quebec on washback at the classroom level and specifically report and discuss the results pertaining to ESL secondary teachers' perspectives when dealing with provincial exams. I conclude with an expanded definition of professionalism, reference to this survey in the larger scheme of educational change, and to the important role that teachers as professionals can play.

Background

An overview of studies demonstrates that the concept of washback is highly complex in nature, contextually bound, and that the stakeholders (e.g., teachers, students, administrators, ministries of education, etc.) appear to be influenced differentially (see Cheng, Watanabe, & Curtis, 2004, for a comprehensive overview; and Alderson & Wall, 1993, for initial hypotheses concerning washback). We are also learning that there are diverse aspects of this phenomenon depending on the sociocultural, sociopolitical, and contextual factors involved, and in addition, depending on the participants involved (Turner, 2001b). We are reminded in the literature that "testing is never a neutral process and always has consequences" (Stobart, 2003, p. 140). Possibly for this reason, the terms *positive* and *negative* have become associated with washback. Bailey (1996) claims that any test (whether "valid" for its purpose or not) can have either positive or negative washback (consequences) depending on whether it enhances or hinders educational innovation and goals.

This brings us to the set of relationships (intended and unintended, positive and negative) across curriculum, teaching and learning, and testing (Fox, 2004). As Pellegrino, Chudowsky, and Glaser (2001) point out, the ideal situation is cohesion across curriculum, instruction, and assessment. This appears easier said than done when one examines educational systems and teachers' behavior and beliefs (Turner, 2002). From a teacher's position, the impact of a high-stakes external test can affect classroom activity in various ways. For example, if such a test represents the curriculum well, and a teacher is teaching the curriculum, then teaching with the general test concepts in mind and preparing students for the test is positive. If a teacher was not focusing on the curriculum, but then became aware of the test content and methodology (which represented the curriculum and innovations in the curriculum), then he or she would ideally change and adjust or align some instruction with general concepts represented in the test. In these situations, elements are synchronized and this would be positive washback. The test results would give the teacher information on student achievement in the program. Therefore, integrating the test's concepts and procedures into the instruction would mean working with students on the abilities they are expected to learn.

On the other hand, if the external test's content and procedures do not represent the curriculum well, then there is a problem (assuming that the teacher is teaching the curriculum). The teacher might abandon the curriculum to prepare the students for an unrelated test. This would be negative washback. In this situation, the test is not serving as an evaluative or assessment tool for the course content. It is not testing what the teacher has been teaching, which is the curriculum. Instead, it is evaluating something else,

which does not give the teacher information on whether the students have achieved or are making progress concerning the curriculum. The ideal situation in an educational system is where the curriculum, teaching, and testing are synchronized, and teachers (and other stakeholders) work for positive washback. Solomon (2002), in her book *The Assessment Bridge*, discusses positive ways to link tests to curriculum improvement.

One must realize that the above is a simplistic explanation of washback at the classroom level. As stated above, we are learning of its complexity. It is important, however, to include in the discussion the voices of teachers who, along with students, are at the grass roots of experiencing test impact at the classroom level. In the past, most reported teacher claims mainly concerned negative washback, for example, narrowing of the curriculum, lost instructional time, reduced emphasis on skills that require complex thinking or problem-solving, and increases in tests scores without a corresponding rise in the ability of the construct being tested (Alderson & Hamp-Lyons, 1996; Andrews, 2004; Barksdale-Ladd & Thomas, 2000; Firestone, Fitz, & Broadfoot, 1999; Linn, 2000). Echoes from the past (Frederiksen, 1984) remind us that efficient tests (e.g., multiple-choice format) tend to drive out less efficient tests (e.g., essays, open-ended interview questions, performance-based tests) leaving important abilities untested and untaught. Many years ago, there were calls for educators and those involved in test construction to develop evaluation instruments that would better represent education goals and to use these instruments to improve the learning process. At present we still see such discussion. Pellegrino et al. (2001), in the book *Knowing What Students Know: The Science and Design of Educational Assessment*, reiterate that educational assessment does not exist in isolation, but must be aligned with curriculum and instruction if it is to support learning. Andrews (2004) states that recently "attention has increasingly been paid to the possibility of turning the apparently powerful effect of tests to advantage, and using it to exert a positive influence in support of curriculum innovation" (p. 39). Currently teachers are trained to conceptualize testing and evaluation procedures as tools to monitor their students' learning. They are encouraged when developing their own in-class instruments to align them with what is being taught. In this way, the assessment procedures serve as a progress or achievement indicator. In this framework, teachers are often also asked to administer high-stakes tests developed externally to their classrooms (e.g., end of year provincial exams). Is there evidence that these high-stakes tests represent the intended curriculum and/or innovations being integrated into the curriculum so that the education system can move ahead in synchronization? In other words, is there evidence of positive washback? (See Pellegrino et al., 2001, for further discussion on revisiting both classroom and high-stakes assessment and how to ensure that both of these approaches together

inform and enhance student achievement.) What is a teacher's professional role in this context?

One way to begin to look at such a question is to seek the perspectives of teachers who are presently working in educational systems with high-stakes exams that are used to assess achievement and to support curriculum innovations. Managing external exams has become a way of life for many teachers. Their ability to deal with them as part of their pedagogical experience is rapidly becoming a professional criterion. Reports from the past describe in general a negative picture of teachers trying to cope. As teachers are trained and become more informed about assessment and the need for synchronization as discussed above, it is important to keep abreast of their perspectives. The rest of this article reports on a teacher survey that is an integral part of a larger study on washback at the secondary level concerning secondary ESL teachers and students in the French school system in the province of Quebec. The teacher survey begins to shed light on a positive washback story as teacher professionalism emerges in dealing with external high-stakes tests.

Methodology

Purpose and Research Questions

The purpose of the teacher survey was to identify the perspectives or beliefs of teachers when a change in the educational system was introduced to them during a school year and then implemented in the end-of-year provincial exam of that same year. In other words, the goal was to explore their views about this situation and the consequences on their behavior and on classroom activity. The major research questions were: What do teachers do in their classrooms when a change in the educational system is introduced through an external high-stakes test? What do they feel is their professional responsibility in reacting to this method to promote curriculum reform? Specifically, the inquiries were to learn about teacher perspectives on how such an innovation affects: what teachers teach (content); how they teach (methodology; e.g., Is it "business as usual" in your classroom? Do you integrate the new ideas into your teaching? Do your classroom teaching content and methodology change? Do your attitudes or beliefs change?).

Population and Context

This study was situated in Quebec, where English-as-a-second-language (ESL) is taught in the school system from grade 3 onward. The participants in the survey were 153 secondary 4 and 5 ESL teachers across Quebec. Detailed information about this sample population is found in *Presentation and Discussion of Results* below.

At the end of high school in Quebec, provincial exams are administered in all mandatory subjects, which include ESL. Theses exams are prepared by teachers and consultants under the MEQ coordination. They are worth 50% of the final mark or grade for students, so students must pass in order to obtain their high school diploma. Under these circumstances these exams are considered high-stakes tests in Quebec.

The Quebec education system is presently undergoing a reform that includes curriculum, organizational, and responsibility changes (Blais & Laurier, 2005). Emphasizing a constructivist approach, the curriculum is competence-based. This is being carried out through a decentralization toward schools and communities and a focus on the importance of teachers' professional judgment and student autonomy. Those involved with the ESL curriculum see this as an opportunity to focus on speaking ability. Classroom instruction and assessment of speaking ability have evolved over the years, but still remain a challenge for many teachers. With French being the first language in the province, exposure to English and the need to speak English are limited (with the exception of sections in the metropolitan area of Montreal). Due to time and resource constraints, some teachers do not focus on practicing and evaluating speaking in the classroom as reflected in the new developing curriculum and goals of the Ministry of Education of Quebec (MEQ) in educational reform. In order to generate more speaking practice, the MEQ decided to introduce specific changes formally (i.e., innovations) into the speaking section of the secondary provincial exam. The intention and hope was that the changes in the exam would be one of several ways to encourage and motivate teachers to practice speaking activities more often with their students and to use English throughout the process.

There were three distinct innovations. The first was a new, empirically derived rating scale that was to be used as speaking assessment criteria (see Gouvernement du Québec, 2004, for the revised scale; Turner, 2001a; Turner & Uphsur, 1996, for the general scale development process). The MEQ put the main emphasis on this new performance rating criteria and the usefulness of it for both teachers and students, as it reflected the curriculum goals for speaking ability. The second innovation introduced was English-only exam instructions (both written and oral) as opposed to instructions in French. The third innovation was a modified speaking assessment task format, that is, a move from one-on-one interviews to student group discussions. In addition, students were allowed individual preparation time before the assessment task.

The speaking assessment task is a group discussion involving three to four students. All students are given instructions in English both orally and written. They are reminded that during the discussion they are to listen to their peers, ask questions, and express their views. Each student chooses a card that has a topic written on it (e.g., decorating my room, the secret to

success, tattooing). They are given a five-minute period to prepare. Students take turns leading a discussion. They are to start by expressing their own views and/or knowledge on the topic. The other students are expected to react by agreeing or disagreeing, asking questions, and so forth. Each student in the group is given a turn to lead a discussion. During this process the teacher circulates and assesses the students individually using the new rating scale criteria.

In order to facilitate the teachers in familiarizing themselves with this new aspect of the curriculum (which was being implemented through the exam), pre-exam actions were taken. Some examples are: groups of teachers were an integral part in developing, validating, and setting standards for the new speaking scale; and workshops, CD-ROMs and written materials on instructional strategies were provided to teachers about the use of instructions in English, speaking group tasks, and how to use the new speaking scale with authentic samples.

Instruments

The instrument used in the survey was a questionnaire composed of two parts (see Appendix). Part 1 asked for background information to help describe the population, and Part 2 asked for teachers' views specifically related to the three innovations introduced into the speaking section of the provincial ESL exam as mentioned above (first innovation, the rating scale—items #2 through #8; second innovation, English-only instructions—items #9 and #10; and third innovation, modified speaking assessment task—items #1 and #11). It also asked general questions on washback beliefs (items #12 and #13). The scale used in Part 2 was a Likert scale ranging from 1=strongly disagree to 4=strongly agree, with the exception of the last three questions, which were open-ended (items #14 and #15 were local procedural questions; and item #16 was a washback-related question about speaking exam preparation). A 4-point scale was purposely used to elicit distinct views and to eliminate the ambiguity of "I don't know" or "I don't have a view." The questionnaire was developed and piloted by the research team for the specific purposes of this study.

Procedure

Participants were recruited through provincial professional forums (i.e., the annual SPEAQ conference, la Société pour la promotion de l'enseignement de l'anglais, language second, au Québec; SPEAQ's interest sections; and SPEAQ's newsletter). Information about the study, including ethical procedures for such survey research, was communicated to the participants. The questionnaire was anonymous and was administered after the provincial speaking exam had taken place; teachers filled it out individually.

Data Analysis

The data from the questionnaire were analyzed using two methods. SPSS 12 was employed for frequency counts (percentages) of the Likert scale questions and for descriptive statistics of the same questions. The added written comments for each question were reviewed to help interpret the numbers. Due to the quantity of comments, an analysis for each question following guidelines as summarized in Tesch's (1990) 10 principles of interpretational analysis was carried out. Comments representing the main patterns are reported in the results. The comments for Q12 and Q13 (i.e., exam affects or should affect teaching and learning), however, were combined with the open-ended question Q16 (i.e., teacher preparation for the speaking exam) for analysis. This was done because the three questions generated comments with much common content and repetition. The qualitative software NUD*IST 4 was used to help organize these data. Due to the overwhelming quantity of responses and to the fact that in general each response was in paragraph form containing several ideas, a qualitative analysis of the comments was carried out similar in nature to open coding as described in Strauss and Corbin (1998) and following guidelines as above (Tesch). Categories were developed from the comments and coded (Bogdan & Biklen, 1998; Marshall & Rossman, 1989). Patterns or themes were identified. I conducted the initial analysis, and another member of the research team, a research assistant, did an independent analysis. Similar categories resulted, but some labeling of the categories differed. Through discussion a consensus was reached as to the wording.

Presentation and Discussion of Results

Participants' Background Details

As stated above, the participants were 153 secondary 4 and 5 ESL teachers in Quebec. Part 1 of the questionnaire revealed that they were all situated in the French school system. Sixty-one percent were female and 39% were male. All participants had BEd degrees, and 3% had MA degrees. All but 4% had had specific ESL training. All but 6% had either taken courses or been involved in workshops on testing and evaluation. There were novice and veteran teachers alike, distributed across four age categories (16% were 20-29 years old; 32% were 30-39; 27% were 40-49; and 25% were over 50). Their teaching experience was distributed across four categories (5% had been teaching for 0-2 years; 16% for 3-6 years; 20% for 7-10 years; and the majority 50% for 11 or more years). The first language for 76% of the teachers was French, for 20% was English, and for 4% was other. The teachers came from nine regions across Quebec with the highest representation coming from Central Quebec (26%) and the Eastern Townships (in southern Quebec) (20%), and the lowest representation from Montreal (6%) and James Bay/Northern Quebec (4%).

Teachers' views relating to innovations introduced into the speaking section of the provincial ESL exam: Perspectives from professionals

As described above, the three new elements implemented in the speaking section of the exam were the rating scale, English (only) instructions, and group discussion tasks with preparation time. Using the definition of washback given above, teachers' perceptions were analyzed to seek evidence of the influence of the new speaking exam components at the classroom level. As we know, teacher views, perceptions, and beliefs are complex constructs. To help gain insight into the data, both the quantitative and qualitative questionnaire data are presented and discussed together so as to provide an interpretative profile. Rather than lengthy descriptions of what teachers wrote, direct quotations represent teachers' voices. The quotations were viewed as representative of the main patterns discovered through the data analysis. Table 1 summarizes the raw data by presenting the percentage of teachers responding in each category on the 4-point Likert scale. For reporting purposes, categories 1 and 2 (strongly disagree/disagree) were combined (i.e., collapsed) into one general category of *disagreement*, and categories 3 and 4 (agree/strongly agree) became one general category of *agreement*. For this study and sample size, the different levels of agreement and disagreement were viewed as being less useful for discussion. Specific levels are only mentioned when pertinent. Table 2 views the data through descriptive statistics.

The teachers agreed that the group discussion format appeared to be an appropriate indicator of students' speaking ability (Q1) and that the new

Table 1
Teachers' Responses: Frequency Counts in Percentages (*n*=153)

Question	*1=S Disagree*	*2= Disagree Agree*	*3=Agree*	*4=S*
1-Exam tasks appropriate indicators	0%	9%	75%	16%
2-Scale accurately measured	2%	9%	64%	25%
3-Felt comfortable using scale	0%	9%	42%	49%
4-Practiced using scale	5%	2%	31%	62%
5-Scale changed my thinking	13%	42%	40%	5%
6-Scale changed my teaching	16%	48%	29%	7%
7-Explained scale to students	5%	2%	37%	56%
8-Students used scale	11%	30%	31%	28%
9-Increased English instructions	33%	17%	24%	26%
10-English instructions problematic	36%	43%	16%	5%
11-Speaking tasks increased	20%	29%	39%	12%
12-Exam affects teaching/learning	2%	26%	61%	11%
13-Exam should affect teaching/learning	11%	34%	46%	9%

Table 2
Teachers' Responses: Descriptive Statistics (n=153)*

Question	*Min.*	*Max.*	*Mean*	*SD*
1-Exam tasks appropriate indicators	2	4	3.07	.50
2-Scale accurately measured	1	4	3.11	.65
3-Felt comfortable using scale	2	4	3.40	.66
4-Practiced using scale	1	4	3.51	.76
5-Scale changed my thinking	1	4	2.36	.77
6-Scale changed my teaching	1	4	2.27	.82
7-Explained scale to students	1	4	3.44	.77
8-Students used scale	1	4	2.76	.99
9-Increased English instructions	1	4	2.43	1.21
10-English instructions problematic	1	4	1.90	.85
11-Speaking tasks increased	1	4	2.44	.95
12-Exam affects teaching/learning	1	4	2.80	.65
13-Exam should affect teaching/learning	1	4	2.52	.82

*Likert Scale: 1=Strongly Disagree, 2=Disagree, 3=Agree, 4=Strongly Agree.

scale helped accurately measure students' ability (Q2). From Table 1, we see a similar pattern in responses. For Q1, 91% of the teachers agreed and for Q2, 89% agreed when categories 3 and 4 are collapsed into one category. The mean and standard deviation in Table 2 indicate this also, 3.07(.50) and 3.11(.65). The combined comments demonstrated teachers' knowledge of the "method effect" (Bachman, 1990), that is, the effects that "task characteristics" (Bachman & Palmer, 1996) including the rating system may have on student performance; and teacher knowledge that student familiarity with the task format may help enhance student performance:

> Good as long as the topic is relevant to the students' reality.
> Good in general, but it depends on other factors, e.g., as long as the students are comfortable with the other group members. It needs to be authentic.
> I like the group format for the exam with everybody talking at the same time because students do not feel like they are being watched and are less shy.
> Next time, I will start using this type of task earlier in the term because it is beneficial to students.
> Some of the topics are too abstract and difficult for the lower spectrum of students.
> [With the scale] it is much easier to give an appropriate evaluation now. We don't have to "guess" anymore.

Teachers indicated that they felt comfortable using the speaking scale (Q3) and took time to practice using it in their classrooms (Q4). When collapsing categories 3 and 4, teacher agreement was 91% and 93% respectively. Table 1 shows that category 4 (strongly agree) obtained the highest percentage of responses for both questions. The means in Table 2 fall approximately in the middle of categories 3 and 4. The comments reveal that teachers took advantage of the information or training sessions that were set up for them and felt confident going back to their classrooms and practicing using the scale.

> Our school board allowed for a complete initiation of how the new scale works. I feel much more comfortable with this scale than the one before. I was able to go back to the classroom and practice a lot and integrate it into my evaluation system.

Teachers who did not have the opportunity to practice (Q4) expressed their frustration in not being able to do so and blamed it on lack of time.

> I had no time, but would like to have more time to let students participate in speaking activities and use the scale.

There was less agreement on whether the new speaking scale changed teachers' ways of thinking about assessment (Q5) and changed their teaching practices (Q6). When collapsing categories 1 (strongly disagree) and 2 (disagree), teacher disagreement with the statements was 55% and 64% respectively. It must be noted, however, that most responses were found in the middle of the scale, categories 2 (disagree) and 3 (agree). This is reflected in the means and standard deviations in Table 2: Q5, 2.36 (.77) and Q6, 2.27(.82). The comments provided insight into the variation of views and also provided teachers' professional stances on their own teaching.

> It [the scale] helped organize my thinking and helped me to mark more fairly.
> It took the "self-interpretation" out of it.
> The scale is a much better tool for assessment and it changed the way I listen to students. I focused more on accuracy and whether the student's discourse was developed and supported. I like the flow chart aspect.
> Didn't really change my thinking. Just that I was now able to more fairly mark.
> Did not change my way of thinking, but I used to modify the old rating scale, placing students between levels; therefore this scale is accommodating. It asks easy-to-answer questions about the students' ability.
> It didn't change my teaching, but it confirmed what I already believed about it.
> Near the end of the year I had to start using it to get the students ready,

for them to get familiar with the approach. So it changed my teaching in some ways. It's a good way to evaluate even though it is not easy.

Teachers reported that they did explain the new scale to their students (Q7) with 93% agreement when collapsing categories 3 and 4 together in Table 1. Table 2 shows a mean and standard deviation of 3.44 (.77). Their comments:

It reassures them [the students] as the final exam seemed to make them a bit more nervous than other oral productions in the year.
Yes, I gave them all a copy.
With the CD provided by the MEQ and a copy of the grid [scale] it was easy to go over it with the students.

There was less agreement on whether students had the opportunity to use the speaking scale themselves (Q8). Table 1 indicates that responses were mainly spread across categories 2, 3, and 4, and Table 2 shows a mean of 2.76 (.99). Comments revealed that teachers' beliefs fell into two areas: (a) students not being able to use such an instrument in that they cannot recognize their own errors; and (b) students should try using the scale to be more aware of what they are being evaluated on.

I don't feel students are ready for that, and they would be too hard on themselves.
I think it's a tool for the teacher, but I explain it to the students.
It is too tough for students because they aren't aware of their own mistakes.
The students all had a copy and had to evaluate themselves. I like to compare theirs to mine.
Yes, it's important, but many of my student's failed to see errors in their own speech.

Q9 and Q10 dealt with the new component of having all instructions on the speaking exam in English. Q9 sought to find out if teachers increased their use of English instructions in the classroom as the exam neared, and Q10 sought views on whether the use of English instructions was problematic. For Q9, Table 1 indicates an even split between agreement and disagreement (50%, 50%) if one combines categories 1 and 2, and then combines 3 and 4, and Table 2 shows a mean and standard deviation of 2.43 (1.21). The comments help make sense of the responses. A portion of the teachers did not increase English instructions because they already conducted their classes in English: "No, because I do everything in English anyway." Others apparently did conduct their classes in French and responded that yes, they did increase exposing their students to English instructions as the exam neared: "Yes English instructions increased in my classroom as the exam approached, but I still sometimes switched to French

to ensure that every student understood and to avoid repeating." At the same time in Q10, although Table 2 indicates that 79% (when combining categories 1 and 2) did not find the use of English instructions on the exam problematic, they expressed concern about the weaker students, but also felt that students in general would now have to make an effort to read the English.

> Yes and No, my students are used to writing tests where all instructions are in English. But the weaker students will miss not having me in the room to explain the odd word in French.
> The weaker students had some difficulty in understanding some parts, not sure of themselves.
> Only my weaker students, because before I would give individual instructions in French for people who were way, way lost.
> I found that the students had always relied on the French translation on the exam. With them now in English, I found they paid more attention to the instructions, therefore, made fewer "stupid" errors.
> No, it was time to do this. Great idea! I always proceeded in English anyway, but I know that in the past exams having instructions in French reduced anxiety levels.

The responses to Q11 (i.e., speaking tasks similar to the exam increased as the exam neared) were nearly evenly split between agreement and disagreement, with 49% in categories 1 and 2 and 51% in categories 3 and 4. Table 2 shows a mean and standard deviation of 2.44 (.95). Once again, the comments provided insight in much the same way as in Q9. A portion of the teachers did not increase such tasks in their classrooms because they were already an integral part of their teaching: "No, because they already have similar activities at regular intervals throughout the year." Another portion who also responded in categories 1 and 2 did not feel they had the time, but it did not matter because they encouraged speaking in English all the time anyway.

> I didn't conduct any special preparation. Nothing like that, no. I was always trying to encourage class discussions or elicit answers in English, but to spend that amount of time on activities similar to the exam, there just isn't time for it.

Those who responded in categories 3 and 4 did increase speaking tasks as the exam approached. Several of the comments revealed that they did this to help the students feel comfortable with the format of the tasks.

> Yes, I increased the tasks despite the fact that the students are in teams all year. I follow procedures a little more, to make them more comfortable.
> I would not have focused so much on the tasks that required students to

talk in groups … if they had not been on the speaking exam. I had to allow students to practice. Overall, I believe this was good.

Teacher views on washback: The impact of the provincial speaking exam

In analyzing the comments from Q12 (i.e., exam affects teaching-learning), Q13 (i.e., exam *should* affect teaching-learning), and the open-ended question Q16 (teacher preparation for the speaking exam), it became apparent that there was much overlap and in reality much repetition. The decision was made to combine all the comments and do a qualitative analysis to reveal patterns or themes in the data (see *Data Analysis*). The teachers had much to say and in many cases connected their comments in the three questions by making cross references. Results from Q12 and Q13 in Tables 1 and 2 show a distinction between teachers' views on whether the exam *does* affect teaching-learning and whether it *should*, with 72% in agreement with the former and 55% in agreement with the latter (when combining categories 3 and 4). The respective means and standard deviations are 2.80 (.65) and 2,52 (.82). In each question, however, the highest percentage is found in category 3 (agree). It was by analyzing the comments related to the two questions, in addition to the comments in the open-ended Q16, that provided a window into teachers' views related to washback.

The categories generated from the three questions are presented in Table 3.

The most salient theme that emerged from the data was teachers' awareness of the importance of the link between teaching and assessment. This recurring theme is articulated in the following comment.

> As a student teacher, I used to believe that evaluation practices should not drive teaching, but now [as a teacher] not only do I realize how much they do, but that the evaluation practices should reflect what students have been taught in the classroom. There is a connection between this and the provincial exam.

Other themes were related to strategies in the classroom in relation to the provincial exam. It appears that teachers felt that aligning classroom practice with the exam construct (i.e., speaking with peers) was important, but they took this alignment to mean different things. Some took it literally and felt

Table 3
Teachers' Perspectives on Washback:
Categories Generated From Q12, Q13, and Q16

- Awareness of link between teaching and assessment
- Aligning teaching practice with exam task characteristics
- Actual classroom strategies
- No special strategies

restricted, whereas others took a broader approach. Some felt obligated to practice the exact format of the task with pertinent vocabulary for expressing opinion and asking questions (i.e., put students in groups and have them take turns leading a discussion). Other teachers felt satisfied that by encouraging students to speak and conducting several speaking activities throughout the year, the students would be sufficiently prepared for the speaking provincial exam. The following comments reflect these various professional stances.

> I believe evaluation procedures should reflect what has been taught in class. So if we stick to the program objectives, then the students should be okay on the final exam … but I sometimes would like to do "different" stuff.
> I agree since evaluation equals the objectives. I disagree since sometimes we miss the point wanting to fill the needs for evaluation.
> I agree, but unfortunately not enough importance is put on formative evaluation. Passing the final exam is not an end in itself. Becoming competent in your second language should be the goal. The whole evaluation system including classroom evaluation should work together.
> I prepare my students throughout the term for the speaking section of the exam by giving them activities similar to the exam.
> Just before the exam, I gave my students a handout with written samples from the teacher's booklet. I also added lists of useful expressions so they had something they could review before the exam.
> We prepared before hand by brainstorming various subjects (through a cooperative approach). We looked at pertinent vocabulary, verb tenses, key expressions for expressing opinions and possible questions that could be asked. This was greatly appreciated by the students.
> I didn't prepare in any special way, because we do regular different speaking activities throughout the whole year, real daily expressions, debates, etc.
> I didn't specifically prepare them other than a previous speaking exam. Throughout the year, they have speaking activities, and the "final" becomes in my opinion the "final activity."

A final theme in the data was teachers' concern in helping and supporting students to perform well. This is naturally related to the strategy themes above, but it went a step farther in illustrating teachers' desire to aid their learners.

> As a teacher, I want my students to do well on the exam so it is important that I prepare them for this.
> Yes, I spent a lot of time preparing my students … I want to make sure they are not surprised when they see the exam. Evaluation means a lot to them. Students often say, "does it count?"

I want my students to speak in English all the time, so that the exam will just be like another conversation, and so that nervousness, mood, etc. will not interfere with their speaking performance.

Throughout the data analysis, as the researcher I was grateful to the teachers for how they provided comments. Their articulate responses aided in interpreting their views.

Discussion

In the context of a larger washback study, the purpose of this survey was to identify the perspectives or beliefs of teachers when a change in the educational system (i.e., ESL speaking practice) was introduced during a school year and then implemented in the end-of-year provincial exam. The inquiry was to examine positive or negative washback as seen through the lenses of teachers. The survey results provide a window into a situation that emerged as much more complex. Rather than simply embracing or rejecting introduced changes, teachers appear to have integrated them into their teaching or assessment practice according to their own beliefs and professional stances. Their reactions seem to reflect that this was all part of a day's work—part of their professional repertoire. We learned that through experience and/or formal education these teachers displayed knowledge about many important elements that abound in the language testing and assessment literature about educational contexts: the effect of method including scoring on student performance (Bachman, 1990, Bachman & Palmer, 1996); the effect of student familiarity with task type and scoring criteria (Genesee & Upshur, 1996; Arter & McTighe, 2001); the importance of linking curriculum, teaching, and assessment (whether the latter be classroom-based or external high-stakes test, Pellegrino et al., 2001; Solomon, 2000); and the understanding that it is at the classroom level where teaching and learning occur and that formative evaluation at this level has an important and different role than high-stakes provincial exams (James & Gipps, 1998). We learned that the teachers believed that they aligned their related classroom practice with the new elements of the provincial speaking exam. This was done in varying ways contingent on the current state of affairs in their respective ESL classrooms. They expressed the belief that the system (both classroom and provincial levels) should work together.

The findings do not indicate what has been the general trend in the literature about teachers' reports, that is, a negative washback story from teachers' perspectives. Instead, in this survey a more positive washback context has emerged. Although some studies have reported teachers' positive attitudes in relation to some aspects of high-stakes tests (Cheng, 2004), and other literature has alluded to this potential (Andrew, 2004) or discussed solutions to create such a context (Solomon, 2000), few studies have reported

on and woven together a profile such as the one in this study. Possibilities for this variation may be attributed to the teachers' stances and perspectives concerning innovations as found in this population of teachers. As Fullan and Stiegelbauer (1991) observe,

> If we know one thing about innovation and reform, it is that it cannot be done successfully to others. It is not as if we have a choice whether to change or not. Demands for change will always be with us in complex societies; the only fruitful way ahead is to carve out our own niche of renewal and build on it. (p. xiv)

Although the innovations in this study were imposed through the provincial exam, they were actually intended curriculum and methodological changes in educational reform (and the MEQ invited teacher participation in aspects of their development). The data demonstrate that the teachers appeared to view them as such and integrated them into their teaching and assessment practice.

Although the results from this survey appear to reflect an image of positive washback, it appears that there is a need to revisit what positive washback might mean in this context. In earlier literature, it is discussed as a pedagogical phenomenon in which the various elements of an educational system (curriculum, teaching, assessment) move toward synchronization when changes are introduced through a high-stakes exam (i.e., innovation theory, Wall, 1999, 2000). This survey has provided insight into the nature of how this might take place from teachers' perspectives. Teachers may or may not embrace the changes, but they cope with them as part of their work and integrate them into their teaching practice. In the process, they express their views as to the nature of the changes. Teachers appear to want to do their part in moving the system into a position where curriculum, their teaching and assessment, and the system's high-stakes exam correspond. They have done this according to their beliefs and professional stances, which in the end may not present a unified performance across teachers. It does, however, demonstrate influence from the final provincial exam on teachers' perceptions of their behavior.

Conclusion

The results of this survey bring us back to the beginning of this article and the discussion about washback, professionalism and innovation theory. The teachers here expressed the will to move ahead with changes that were introduced through the speaking exam and to move toward a synchronization of curriculum, teaching, and assessment in general. In this professional stance, it became apparent, however, that they struggled at times with factors pointing toward a need for better alignment between assessments used for different purposes (classroom-based assessment and high-stakes provin-

cial exams). The teacher perspectives and stances that emerged contribute to the multifaceted concept of professionalism.

With enhanced student learning as the intended goal, more efforts and research are needed in order that assessments at all levels work together in a system that is comprehensive, coherent, and continual (Pellegrino et al., 2001). The role that teachers can play at the classroom level is revealed in the professional stances that emerged. This is yet another indication of the pivotal role that teachers can play in our educational contexts.

Acknowledgments

This project was funded by a grant from the Social Sciences and Humanities Research Council (SSHRC), and in addition by SPEAQ (la Société pour la promotion de l'enseignement de l'anglais, language second, au Québec). I thank my research assistants, Yvonne Christiansen, Christian Colby-Kelly, Tim Dougherty, Kerry Hatsipantos, and Christopher Sikorsky; and Catherine MacDonald and Elizabeth Johnston at the MEQ for their assistance. I also express my appreciation to all the contributing teachers who took the time to provide rich data. Also, thanks to the anonymous reviewers for their useful comments.

The Author

Carolyn E. Turner is an associate professor and Director of Graduate Programs in the Department of Integrated Studies in Education at McGill University. Her main focus and commitment are language assessment and testing in educational settings. She pursues these through her teaching, research, and service. She has published in journals such as *Language Testing, TESOL Quarterly*, and the *Canadian Modern Language Review*. She is currently Associate Editor of *Language Assessment Quarterly.*

References

Alderson, J.C., & L. Hamp-Lyons. (1996). TOEFL preparation courses: A study of washback. *Language Testing 13*, 280-297.

Alderson, J.C., & Wall, D. (1993). Does washback exist? *Applied Linguistics 14*, 115-129.

Andrews, S. (2004). Washback and curriculum innovation. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 37-50). Mahwah, NJ: Erlbaum.

Arter, J., & McTighe, J. (2001). *Scoring rubrics in the classroom*. Thousand Oaks, CA: Corwin Press.

Bachman, L.F. (1990). *Fundamental considerations in language testing*. Oxford, UK: Oxford University Press.

Bachman, L.F., & Palmer, A.S. (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.

Bailey, K.M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing, 13*, 257-279.

Barksdale-Ladd, M.A., & Thomas, K.F. (2000). What's at stake in high-stakes testing: Teachers and parents speak out. *Journal of Teacher Education, 51*, 384-397.

Blais, J.G., & Laurier, M. (2005, April). *Accountability and standardized testing in Quebec and some neighbouring US states*. Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Bogdan, R., & Biklen, S. (1998). *Qualitative research for education*. Cambridge, UK: Cambridge Unversity Press.

Cheng, L. (2004). The washback effect of a public examination change on teachers' perceptions toward their classroom teaching. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and method* (pp. 147-170). Mahwah, NJ: Erlbaum.
Cheng, L., Watanabe, Y., & Curtis, A. (Eds.). (2004). *Washback in language testing: Research contexts and methods*. Mahwah, NJ: Erlbaum.
Englund, T. (1996). Are professional teachers a good thing? In I.F. Goodson & A. Hargreaves (Eds.), *Teachers' professional lives* (pp. 75-87). London: Falmer Press.
Firestone, W.A., Fitz, J., & Broadfoot, P. (1999). Power, learning and legitimation: Assessment implementation across levels in the United States and the United Kingdom. *American Educational Research Journal, 36*, 759-793.
Fox, J. (2004, October). *Language test impact: Practices and possibilities*. Paper presented at the meeting of the Midwestern Association of Language Testers, Cleveland.
Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist 39*, 193-202.
Fullan, M.G., & Stiegelbauer, S. (1991). *The new meaning of educational change* (2nd ed.). New York: Teachers College Press.
Genesee, F., & Upshur, J.A. (1996). *Classroom-based evaluation in second language education*. Cambridge, UK; New York: Cambridge University Press.
Gouvernement du Québec, Ministère de l'Éducation. (2004). *Document d'Information: Épreuves Uniques, Anglais, Langue Seconde, de quatrième et cinquième année de secondaire 156-444 et 156-544* (No. 16-7105-05, 16-7181-05). Retrieved October 15, 2004, from http://www.meq.gouv.qc.ca/dgfj/de/docinfosec.htm
Hedgcock, J.S. (2002). Toward a socioliterate approach to second language teacher education. *Modern Language Journal, 86*, 299-317.
Henrichsen, L.E. (1989). *Diffusion of innovations in English language teaching: The ELEC effort in Japan*. New York: Greenwood Press.
James, M., & Gipps, C. (1998). Broadening the basis of assessment to prevent the narrowing of learning. *Curriculum Journal, 9*, 285-297.
Kumaravadivelu, B. (2003). *Beyond methods: Macrostrategies for language*. New Haven, CT: Yale University Press.
Linn, R.L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4-16.
Marshall, C., & Rossman , G.B. (1989). *Designing qualitative research*. Newbury Park, CA: Sage.
Mathews, P., & Chuntian, C. (Eds.). (2004). Professionalism in teaching English as a second language (ESL) in Canada and abroad. *TESL Canada Journal, 4*, special issue No. 4, i-ii.
McNamara, T. (1998). Policy and social considerations in language assessment. *Annual Review of Applied Linguistics, 18*, 304-309.
Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*, 241-256.
Pellegrino, J.W, Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.
Solomon, P.G. (2002). *The assessment bridge: Positive ways to link tests to learning, standards, and curriculum improvement*. Thousand Oaks, CA: Corwin Press.
Stobart, G. (2003). The impact of assessment: Intended and unintended consequences. *Assessment in Education, 16*(2), 139-140.
Strauss, A., & Corbin, J. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (2nd ed.). Thousand Oaks: Sage.
Tesch, R. (1990). *Qualitative research: Analysis types and software tools.* New York: Falmer.
Turner, C.E. (2001a). *Developing an empirically based rating scale for evaluating speaking ability at the secondary 4 and 5 levels* (Report). Quebec: Ministry of Education.
Turner, C.E. (2001b). The need for impact studies of L2 performance testing and rating: Identifying areas of potential consequences at all levels of the testing cycle. In A. Brown, C. Elder, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara, & K O'Loughlin (Eds.),

Experimenting with uncertainty: Language testing essays in honour of Alan Davies, Studies in Language Testing 11 (pp. 127-139). Cambridge, UK: Cambridge University Press.
Turner, C.E. (2002, December). *Investigating high-stakes test impact at the classroom level*. Paper presented at the annual meeting of the Language Testing Colloquium, Hong Kong.
Turner, C.E. (2005, May). *Professionalism and the impact of high-stakes tests at the classroom level: The speaking component of the ESL Provincial Exam in Quebec*. Paper presented at the TESL Canada Conference, Ottawa.
Turner, C.E., & Upshur, J.A. (1996). Developing rating scales for the assessment of second language performance. In G. Wigglesworth & C. Elder (Eds.), *Australian review of applied linguistics: Series S, No. 13. The language testing cycle: From inception to washback* (pp. 55-79). Melbourne: ARAL.
Wall, D. (1999). *The impact of high-stakes examinations on classroom teaching: A case using insights from testing and innovation theory*. Unpublished doctoral dissertation, University of Lancaster, UK.
Wall, D. (2000). The impact of high-stakes testing on teaching and learning: Can this be predicted or controlled? *System, 28*, 499-509.

Appendix

(TURNER, 2004, WB PROJECT)

Final Teacher Questionnaire

This information will help us understand better your impressions of the speaking section of the final provincial examination and its relation to teaching activities. All information will be treated in the strictest confidence. Thank you very much for your time.

Part 1: Your Background Information

Please check [ ] the appropriate answer.
(1) Your gender: [ ] male [ ] female
(2) Your age: [ ] 20-29 [ ] 30-39 [ ] 40-49 [ ] above 50
(3) Your mother tongue: [ ] English [ ] French [ ] Other—Specify _____
(4) Number of years you have been teaching: [ ] 0-2 years [ ] 3-6 years [ ] 7-10 years
[ ] 11 years or more
(5) Number of hours you teach ESL per week:
[ ] 0-4 hours [ ] 4-10 hours [ ] 11 hours or more
(5a) Levels you teach: [ ] Secondary 4 [ ] Secondary 5 [ ] Other, Specify
(5b) Class types: [ ] Regular ESL [ ] ESLA [ ] Enriched [ ] Other, Specify
Comments on (5, 5a, 5b):
(6) Your academic background: [ ] Bachelors [ ] Bachelors plus Certificate [ ]Masters
[ ] PhD[ ] other, Specify: _____
(7) Do you have specific training in ESL? [ ] Yes [ ] No
(8) Have you taken courses specifically in testing and evaluation? [ ] Yes [ ] No
(9) Have you been involved in workshops focusing on testing/evaluation?
[ ]Yes [ ] No
(10) Region you teach in:
[ ] Montreal [ ] Montreal region (Laval, South Shore) [ ] Eastern Townships
[ ] Laurentians

[ ] Quebec City [ ] Central Quebec (Mauricie, Charlevoix, Chaudiere regions)
[ ] Saguenay-Lac-St-Jean region [ ] Western Quebec and Hull region
[ ] Eastern Quebec (Gaspe region, Manicouagan, Duplessis and the Magdalen Islands)
[ ] James Bay and Northern Quebec Specify city/municipality: _____

Part 2: Speaking Evaluation

In the brackets [ ], please mark the following on a four point scale as:
[1] strongly disagree [2] disagree [3] agree [4] strongly agree

(1) [ ] I believe the **speaking activities** on the final exam are an appropriate indicator of the student's ability.
Comments:

(2) [ ] I believe the ***new*** **speaking scale** for the final provincial examination accurately measured the speaking ability of my students.
Comments:

(3) [ ] I felt comfortable using the ***new*** **speaking scale** in the final provincial examination.
Comments:

(4) [ ] I had the opportunity to practice using the ***new*** **speaking scale** before final provincial speaking evaluation.
Comments:

(5) [ ] The ***new*** **speaking scale** changed my way of thinking about the assessment of my students.
Comments:

(6) [ ] The ***new*** **speaking scale** changed my teaching in some ways.
Comments:

(7) [ ] I had the opportunity to explain the ***new*** **speaking scale** to my students.
Comments:

(8) [ ] I had the opportunity to have my students use the ***new*** **speaking scale** themselves.
Comments:

(9) [ ] The amount and frequency of **English instructions** increased in my classroom as the final examination approached.
Comments:

(10) [ ] I felt having the **instructions in English** in the final provincial examination to be problematic for my students.
Comments:

(11) [ ] The amount and frequency of speaking tasks similar to the final speaking examination increased in my classroom as this examination approached.
Comments:

(12) [ ] I believe evaluation procedures drive (affect) teaching/learning.
Comments:

(13) [ ] I believe evaluation procedures ***should*** drive (affect) teaching/learning.

Comments:

Please answer the following questions in your own words.

(14) How many weeks after you received it did you administer the *speaking section* of the provincial examination?

(15) What factors affected this timing?

(16) Please comment below on whether you prepared your students for the speaking section of the provincial exam. If you did, please comment on how you prepared your students.