Unintended Consequences of using Tests to Improve Learning: How Improvement-Oriented Resources Heighten Conceptions of Assessment as School Accountability

Gavin T. L. Brown
Hong Kong Institute of Education

Lois R. Harris
University of Auckland

ABSTRACT: Over the past decade, the New Zealand government has created a set of resources to support teachers' use of assessment *for* learning. These include Assessment Tools for Teaching and Learning (asTTle), a software program enabling teachers to create personalized but standardized tests for diagnostic purposes, and Assess to Learn (AtoL), an intensive professional development program. These resources were expected to increase teacher agreement that improvement is the major purpose of assessment. Instead, a 2008 sample of teachers completing the Teacher Conceptions of Assessment questionnaire showed significantly higher agreement that assessment is about school accountability than participants in previous national surveys. Unlike previous surveys, the correlation between school accountability and improvement conceptions was not statistically significant. However, as only the improvement conception predicted the practices teachers used to define assessment (β = .32), it appears that these teachers still saw many of the practices they used in the classroom (e.g. oral and interactive assessments) as improvement-oriented. Interviews with twenty-six of the surveyed teachers identified that while a few saw the new resources as contributing to improvement and accountability purposes, a larger group failed to make that connection. This second group seemed to be unable to accept that tests, an assessment genre traditionally associated with school and student accountability, could be meaningfully used for improvement at the classroom level. These data show that schools and individuals mediate the implementation of any policy initiative and can therefore cause it to have a range of often unintended consequences. In light of this, the thinking of teachers and other educational stakeholders should be taken into account when enacting policy changes.

KEYWORDS: *teacher thinking; teacher attitudes; conceptions of assessment; school accountability/evaluation; elementary and secondary teachers; standardized tests; professional development*

Teachers' opinions, attitudes, and beliefs (a.k.a., conceptions, Thompson, 1992) play an important part in mediating how educational reforms are implemented in schools and classrooms (Richardson & Placier, 2001).

This article examines and interprets observed changes in teacher thinking about the purpose of assessment following the New Zealand government's provision of assessment resources designed to improve student learning. Explicit attention to teachers' conceptions related to the purposes of assessment and how those conceptions influence their practices is used to shed light on the value and worth of the policy resources. Insights from the teachers, much in the spirit of quality control circles, is used to identify factors that are impacting on the success of these initiatives to promote assessment directed at improving student learning. To provide a background for this study, literature on teacher conceptions is reviewed prior to descriptions of the New Zealand context of this study.

Teacher Conceptions

Ajzen's (2005) model of planned or reasoned behavior suggests that teachers' intentions and beliefs about what others think and their sense of power to fulfill these intentions determine their behaviors within school environments. Conceptions capture what teachers think about the nature and purpose of educational processes and practices (Thompson, 1992) and often originate from their own personal educational experiences (Pajares, 1992). Evidence exists that teachers' conceptions of various educational processes (e.g., teaching, learning, and curricula) strongly influence how they teach and what students learn and achieve (Clark & Peterson, 1986; Thompson, 1992; Calderhead, 1996). This study particularly draws on the literature showing that teachers' beliefs about students, learning, teaching, and subjects influence their assessment techniques and practices (Asch, 1976; Cizek, Fitzgerald, Shawn, & Rachor, 1995; Kahn, 2000; Tittle, 1994).

Educational policy shapes the context in which teachers perform their multifaceted work (e.g. planning, teaching, and evaluating). Policy expresses the societal and cultural norms valued by members of that area, most of whom are not teachers. Thus, the introduction of policy reform around assessment (e.g., No Child Left Behind) may express values not necessarily held by those employed to implement the policy (i.e., teachers). Hence, attention needs to be paid to the conceptions teachers have surrounding current practices and appreciate how they are most likely to understand, respond to, and implement reforms.

While research has examined teachers' assessment and grading practices (e.g., McMillan, 2001; McMillan, Myran, & Workman, 2002; Stecher & Barron, 2001), the majority of those investigations have not focused on the purposes or intentions teachers have for these practices. It may be that certain practices are seen as improving student learning; however, within different contexts, the same practices could be viewed as fulfilling administrative or accountability goals. For example, in a study of New Zealand primary school teachers that used items from McMillan's (2001) questionnaire, Brown (2009) showed that two different conceptions of assessment (i.e., assessment improves learning and assessment is irrelevant) equally predicted the use of interactive-informal assessment practices. Hence, explicit attention to teachers' thinking about the purposes of assessment and related practices such as reporting and grading is needed to better understand how policy changes actual practice.

As conceptions are a function of embodied experiences (Lakoff & Johnson, 1999), it is expected that differences in culture or society lead not only to differing policies, but also distinct conceptions of practices or processes. For example, Hamilton et al. (2007) reported that teachers in California, Georgia, and Pennsylvania had very similar responses,

experiences, and attitudes towards standards-based accountability assessments; they attributed this to similarities between the systems. In contrast, teachers in New Zealand and Hong Kong had very different understandings of how the practice of grading students related to improved learning (Brown, Kennedy, Fok, Chan, & Yu, in press). In Hong Kong, agreement with the conception that assessment evaluates students was very strongly and positively correlated (r = .91) with the conception of assessment for improvement; in New Zealand, the same two conceptions were very weakly correlated (r = .21). The difference was attributed to cultural features of the Confucian system in Hong Kong that emphasizes educational testing as a force for improved learning.

The model underlying the research reported in this paper (see Figure 1) is loosely similar to Hamilton et al.'s (2007) conceptual framework. The model used in this study has twin, interacting tracks leading to student outcomes; within it, the conceptions of both teachers and students are influenced by various policy directions and family priorities and these beliefs, in turn, guide their separate teaching and learning practices. These two pathways are shaped by and respond to societal and cultural contexts, meaning that there will be different beliefs and practices in differing social, ethnic, and cultural groups. Note that this model does not attempt to portray the complex paths leading to teachers' and students' conceptions, which have been hinted at in Pajares (1992). There are three important distinctions between this model and Hamilton et al.'s (2007). First, teacher beliefs are seen as mediating between policy and outcomes, rather than as external to the implementation processes. Second, policy directions are seen as a function of priorities within society and culture, suggesting that variation in conceptions and practices within societal contexts will be less than those between contexts. Third, students themselves are thought to have a strong contributing role in shaping their outcomes.

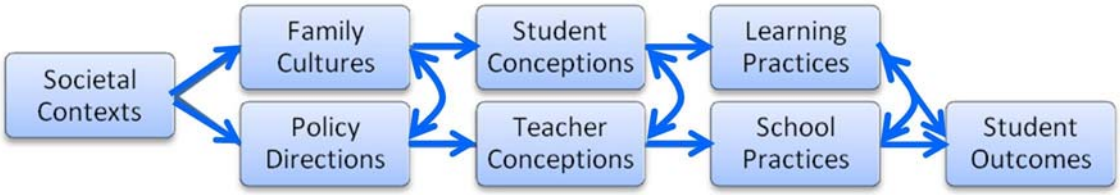

Figure 1. Conceptual Framework of Relations Leading to Outcomes

Conceptions of Assessment

One of the difficulties in researching teachers' conceptions of assessment is that they appear to hold multiple and, at times, contradictory conceptions without being disturbed by such contradiction (Cizek, Fitzgerald, Shawn, & Rachor, 1995; Kahn, 2000). This is perhaps because assessment serves multiple purposes. Four conceptions of assessment exist, which may loosely be categorized as three purposes and one "anti-purpose" (Brown, 2008). The three major purposes for assessment are improving teaching and learning, making schools and teachers accountable for their effectiveness, and making students accountable for learning (Heaton, 1975; Torrance & Pryor, 1998; Warren & Nisbet, 1999; Webb, 1992). An anti-purpose can be detected in practices that treat assessment as fundamentally irrelevant to the life and work of teachers and students (Shohamy, 2001).

The improvement purpose, which involves both teachers and students using assessment to improve either teaching or learning, has been shown to have positive impacts on educational outcomes (Black & Wiliam, 1998; Crooks, 1988; Popham, 2000). The school accountability purpose uses student assessment results to establish the quality and effectiveness of school practices and is often associated with high-stakes consequences for schools and teachers (e.g., Hershberg, 2002; Linn, 2000; Noble & Smith, 1994; Smith & Fey, 2000). The central tenet of student accountability is that assessments are used to either motivate students or publicly certify the quality of learning students have achieved (Guthrie, 2002). The final conception is based on a rejection of external evaluation processes as being inadequate, inaccurate, and/or irrelevant to the teachers' ability to improve student learning (Shohamy, 2001).

Assuming that these four conceptions are exclusive and exhaustive and that there are three fundamental stances towards each conception (i.e., positive, neutral, and negative), there are eighty-one logically different profiles possible. Furthermore, because the conceptual space involves two major dimensions (i.e., student versus school and improvement versus irrelevance) (Harris & Brown, in press), there are complex interrelationships in teachers' understanding of the nature and purpose of assessment. It is no wonder, then, that teachers have complex, idiosyncratic conceptions of assessment.

Policy Effects on Conceptions of Assessment

Differing cultural and policy contexts have already been shown to affect the way teachers' accept and reject the four purposes of assessment discussed in the previous section (Brown et al., in press). Logically, it would be expected that a policy that prioritizes the educational improvement purpose for assessment, as in the case of New Zealand (Ministry of Education, 1994), would lead to teachers who were strongly committed to the notion of assessment for improvement (Brown, 2004b). In contrast, the imposition of high stakes consequences in response to national testing in the United States since the late 1980s has generated much antipathy towards assessment. Teachers regularly attribute undesirable effects such as reduced professionalism, restricted teaching practices, and narrowed student learning outcomes to inappropriate external testing (Darling-Hammond, 2003; Hamilton, 2003; Hamilton et al., 2007; Linn, 2000). However, it is now being reported that not only are positive consequences from national testing occurring (e.g., Black & Wiliam, 2004; Cizek, 2001; Monfils et al., 2004), but that teachers are also aware that accountability pressures can and do lead to educational improvement (Hamilton et al., 2007). Hence, it would seem that the relationship between school accountability and improvement conceptions in the minds of teachers are not simple opposites.

Research with New Zealand teachers has shown that within the context of a policy framework that emphasizes assessment *for* improvement, with responsibilities at the school level for monitoring and reporting progress to parents and government, teachers emphasized both the improvement and student accountability conceptions of assessment (Brown, 2008). However, the correlations between the two accountability conceptions and improvement were intriguing. Whereas the teachers rejected the notion that assessment was about making schools accountable, there was a moderate positive correlation between improvement and school accountability (r = .46). In contrast, there was moderate positive correlation between student accountability and irrelevance (r = .36). Furthermore, it was shown that the school accountability conception predicted the use of deep learning assessment practices, while the student accountability conception predicted the use of surface learning

and test like assessment practices (Brown, 2009). Because improvement was prioritized by the low stakes assessment policy framework of New Zealand, together these patterns suggest that teachers wanted to use assessment to demonstrate school quality, but believed that assessment systems must measure highly valued outcomes, such as deep learning. Consequently, in the interim, they were cautious about the power of external, test-like assessments to evaluate schools fairly.

Hence, there are both empirical and theoretical reasons for believing that changes in policy framework (e.g., the introduction of a reform) would lead to differences in how assessment is conceived. It is presumed that a low-stakes assessment environment would encourage greater adoption of improvement rather than irrelevance conceptions. It is also presumed that high-stakes consequences for schools and teachers would lead to a weak association between improvement and school accountability conceptions and a greater focus on students as the people being held accountable through assessment.

Research Context

The New Zealand Ministry of Education is committed to the use of assessment *for* learning, expecting schools to use assessment evidence for improving student learning. In the last decade, two significant resource initiatives have been deployed nationally to assist teachers with assessment practices (Brown, Irving, & Keegan, 2008; Crooks, 2002). The first is the progressive release of national computer-assisted assessment tools from 2002 onwards (i.e., Assessment Tools for Teaching and Learning—asTTle; Hattie et al., 2004). The second is the national provision of assessment-focused professional development services since 2003 (i.e., Assess to Learn—AtoL). It is important to realize that there are no compulsory national assessments or tests in the New Zealand school sector (Crooks, 2002). System monitoring takes place in Years 4 and 8 through the National Education Monitoring Project light sampling of student performance, while schools may use any of a range of standardized tests or informal procedures to diagnose student learning needs (Brown et al., 2008). The high-stakes assessment of students takes place in the final three years of secondary schooling (Years 11 to 13) through a combination of internally-administered assessments and external end-of-year examinations. School quality is determined through triennial reviews by the Education Review Office, which does not require that schools demonstrate effectiveness with any one assessment method. Hence, tests and examinations in New Zealand are evaluative for students (especially in the final years of schooling); whereas, standardized tests function, for schools, as improvement-oriented assessments. Thus, it was a legitimate expectation that a new test system could be seen as an adjunct to assessment for learning rather than as an evaluative accountability mechanism.

asTTle in NZ

Costing more than NZ$17 million, the Assessment Tools for Teaching and Learning (asTTle) software is the single most expensive New Zealand policy-based assessment tool. asTTle is a national curriculum- and normative-referenced educational resource that makes use of advanced computer technology. The development and use of asTTle has been seen as a solution to the negative effects of compulsory national testing, while meeting accountability requirements (Hattie & Brown, 2008). This system has been described extensively elsewhere (Brown, Irving, & Keegan, 2008; Crooks, 2002; Hattie & Brown, 2008; Hattie, Brown, & Keegan, 2003), so only a brief overview is given here.

asTTle provides the autonomous, decentralized schools of New Zealand an educational technology resource that permits both improvement and reporting responses to

assessed student performance in reading, writing, and mathematics in either English or Maori. Since 2002, schools have been provided the asTTle software free of charge; usage is completely voluntary. The asTTle software allows schools and teachers to create curriculum-aligned customized, standardized, forty-minute mathematics, reading, and writing tests from large banks of calibrated questions. This level of personalization was designed to allow teachers to test students at their own levels, ensuring test results were meaningfully related to what students were learning and could provide teachers and students with useful feedback on pupil progress. Reporting is against both the objectives and strands of Curriculum Levels 2 to 6 and norms for students in Years 4 to 12. All asTTle items and tasks were mapped by teachers, content area experts, and curriculum experts according to the NZ curriculum statements for the relevant subjects. Additionally, all items were mapped to a cognitive processing taxonomy (i.e., the Structure of Observed Learning Outcomes—SOLO) in order to categorize student performance on the various tasks according to broad levels of current functioning (Hattie & Brown, 2004). The test-users can select from a suite of graphical reports (including an online catalogue of curriculum-aligned teaching resources to help teachers respond appropriately to assessment data) that allow interpretation of the performance of individuals and cohorts relative to norms, standards, and objectives; these were designed to suit both improvement and accountability purposes (Hattie, Brown, Ward, Irving, & Keegan, 2006). asTTle's graphic reports were designed so student results could be shared with parents and pupils and thereby used to involve them in the diagnosis and educational goal setting integral to an assessment *for* learning approach. At the time of this study, schools were using asTTle Version 4 (Hattie et al., 2004).

While the explicit focus in the asTTle design is on assessment for improved learning and teaching, it is possible, through the various reporting systems, to use the data to demonstrate school accountability. Since the data are externally referenced to both year norms and curriculum levels, it is possible that teachers could associate it with school accountability mechanisms (i.e., a proxy for national testing) more than assessment for improvement.

Assessment Professional Development

New Zealand has a history of government-funded professional development programs. Between 1995 and 2001, schools could participate in Assessment for Better Learning (ABeL). Each year, approximately 400 schools began the two-year program (Education Gazette, 2002), designed to increase teachers' assessment literacy, knowledge, and application of formative or improvement-oriented assessment practices (Ministry of Education, 2001). A formal evaluation found that the program was associated with substantial, beneficial effects in most participating schools (Peddie, 2000). Such results included better schoolwide assessments, changes in teacher thinking, greater understanding of assessment, and improved reporting. The current Assess to Learn (AtoL) service replaced ABeL and is provided, competitively with budgeted restrictions, to schools (i.e., not all applicants enter the program) in order to improve schoolwide selection, administration, interpretation, and responses to assessment data. For example, the program is designed to help teachers use assessment data (including asTTle test results) to create an assessment *for* learning school environment. Teachers are taught, in accordance with the school's own needs and priorities, the logic and practice of sharing learning intentions and success criteria with students and a range of ways to provide students with meaningful feedback about their learning. AtoL providers are regionally-based, with contracts competitively awarded by the

Ministry of Education. All authorized providers are expected to support the government's assessment policy and resources, but are free to deliver their service according to their own understanding of effective and valid assessment usage. Only 200 schools nationally received the AtoL service between 2005 and 2007 (Poskitt & Taylor, 2008).

One of the current emphases in the AtoL program is supporting school-wide data usage for school improvement; student assessment data are analyzed by school leadership teams to identify teaching and learning priorities. Since the goal of school improvement is what it says—improved learning—it is possible such an emphasis would align well with the conception that assessment is for improvement. However, it is also equally possible that teachers would conceive of this as a means of monitoring and accountability since school leaders normally undertake schoolwide analysis. Furthermore, since the drive for school improvement is often seen as an external force, it is possible that this priority could lead to a greater conceptualization of assessment as a school accountability mechanism, rather than improvement.

Method

Study Design

The study reported in this paper was part of the Measuring Teachers' Assessment Practices (MTAP) project at The University of Auckland. The goal of the MTAP project is to explore the relationships among teachers' conceptions of assessment, teachers' assessment practices, students' conceptions of assessment, and students' academic outcomes. MTAP Study 1, reported here, was a replication, in part, of a previous nonexperimental survey that took place in 2001 (Brown, 2002), before the release of asTTle and AtoL. Given these reforms, it was decided to examine a contemporary sample of teachers in the Auckland region to ascertain whether these programs were having an impact on teachers' conceptions of assessment. The analysis reported in this paper was a secondary analysis of MTAP Study 1 data. This analysis was conducted to see if any conceptual shifts occurred since the implementation of asTTle and AtoL and to identify what role, if any, teachers said these tools had in influencing their thinking and practices surrounding assessment. It was expected that the implementation of asTTle and AtoL (both intended to lead to greater use of assessment for improved teaching and learning) would lead to greater teacher agreement with the notion that assessment should be for student improvement than had been found in previous surveys. In order to gather more in-depth data on participant thinking, a sample of teachers with diverse response patterns to the questionnaire took part in a semistructured interview about their conceptions of assessment.

While New Zealand has completed a major revision of its curriculum framework in the same time period (Ministry of Education, 2007), this study focused specifically on assessment-related resources and policies that might be expected to influence teachers' conceptions of and practices of assessment. Since this study uses a nonexperimental design and *post hoc* analysis of volunteers to investigate reasons for the observed phenomena, the validity of the interpretations is specifically threatened by history (i.e., a major curriculum revision was launched in the year before this study), selection (i.e., both schools and teachers had to volunteer for the survey), and mortality (i.e., about 40 percent of the survey participants were unwilling to be interviewed). Furthermore, the study is implicitly an evaluation of the two assessment reform initiatives; the interview data and analysis permit the identification of explanations related to how the reforms are actually implemented in a sample of schools for changes in teachers' thinking.

Participants

Thirty-six schools in the Auckland region participated in this study. These included primary (Years 1-6), full primary (Years 1-8), intermediate (Years 7 and 8), and high schools (Years 9-13) from across school deciles (i.e., an index of socio-economic status with 1 being the lowest and 10 being highest). Teachers of Year 5 to 10 mathematics and/or English at these schools were invited to complete Brown's TCoA-III questionnaire (Brown, 2006). In addition, they provided demographic information about themselves and selected from a list the assessment practices they associated with the term "assessment."

Of the 425 questionnaires distributed, 161 were returned (response rate = 38%). Of those returned, 100 (62%) indicated willingness to be interviewed. Profiles were created by examining each teacher's conceptions score and classifying it as within, above, or below the previously established national sample means. After the first author analyzed the questionnaire results, he selected twenty-six teachers with noticeably different conception profiles for interview (Brown & Harris, 2008 provides details of questionnaire results). These teachers then participated in a semistructured interview with the second author.

Instruments

Survey. The instrument used in this research was the abridged, twenty-seven-item *Conceptions of Assessment Inventory* (CoA-IIIA) (Brown, 2006) designed to elicit teacher self-ratings for four main conceptions of assessment (i.e., assessment improves learning, assessment makes schools accountable, assessment makes students accountable, and assessment is irrelevant). The factor structure is hierarchical with nine first-order factors that are, in turn, predicted by the four intercorrelated conceptions. The validity of the instrument was established in a series of studies with New Zealand primary teachers (Brown, 2002). This model, based on New Zealand teachers' responses to the TCoA inventory, had acceptable psychometric characteristics (χ^2 = 841.02; df = 311; χ^2/df = 2.70, p = .10; gamma hat = .93; RMSEA = .057).

Teachers indicate their level of agreement to each statement using a positively-packed agreement rating scale, that is, there were two negative (i.e., mostly and strongly disagree) options and four degrees of positive agreement (i.e., slightly, moderately, mostly, and strongly agree) (Brown, 2004a). Such a skewed response scale has been found useful when participants are likely to agree with statements because the greater range of options within the generally positive range elicits greater variation in responses than when only two response points are used to capture positive orientation.

As assessment is associated with a wide variety of practices, one way to understand how teachers conceive of assessment is to identify the types of assessment they have in mind while completing the questionnaire. Just prior to the CoA-III questionnaire itself, teachers were asked to identify which of up to twelve different assessment practices (i.e., unplanned observation, oral question and answer, planned observation, student written work, group or individual projects, student self or peer assessment, conferencing, portfolio/scrapbook, teacher made written test, standardized test, essay test, and one-to-three hour examination) they had in mind when they thought of the word assessment. Each item was scored dichotomously (i.e., 0 = not selected, 1 = selected) and teachers could choose up to 12. Additionally, teachers were given two blank spaces entitled "other" where they could write in additional practices not listed. Eleven of the items were originally used in Brown (2002) and multidimensional analysis found four clusters of practices (i.e., oral, examination, teacher-controlled, and portfolio).

Interview. The second author carried out semistructured interviews about assessment and its purposes in order to understand the belief systems of participants, without prior knowledge of their questionnaire responses. In hour-long semistructured interviews held at their places of work, participants were asked about their conceptions of the nature and purpose of assessment. These questions included ones such as these:

- Give an example of an assessment activity you used recently in your classroom.
- Describe the purposes of the assessment activity you just described.
- What do you think is the best way to assess student learning?
- To what extent do you believe your personal beliefs about the purpose of assessment align with those promoted by your year group/department? By your school? By government policies and practices?

The last prompt in this list was the one used to obtain data about teacher experiences with AtoL (if they had participated in it) as teachers generally talked about their professional development experiences when comparing their own beliefs to those promoted by the government and their school. Teachers were also directly asked what formal assessment systems they used with students (e.g. asTTle, Progress and Achievement Tests) and encouraged to explain how and why these were used. All data were audio-recorded so they could be transcribed verbatim.

Analysis

Survey. Structural equation modeling was used to analyze the questionnaire and assessment definitions data. First, the original, hierarchical, intercorrelated model was tested for fit to the complete sample of questionnaire respondents, regardless of interview status, to reduce instability in estimation due to small sample size. Second, a measurement model of the assessment type list was developed with exploratory factor analysis and restrictive factor analysis. Finally, a structural model linking the four major conceptions of assessment to the assessment types was explored; only statistically significant paths were retained.

The quality of fit for the specified model to the underlying data matrix is statistically tested with a number of effective measures (i.e., those least affected by sample size). In the studies reported here, acceptable fit is imputed when root mean square error of approximation (RMSEA) is < .08, even if comparative fit index (CFI) is < .90. Conventionally, goodness-of-fit indices (e.g., CFI and gamma hat) should be greater than .95, while badness-of-fit indices (e.g., RMSEA and standardized root mean residual [SRMR]) should simultaneously be ≤ .05 (Hoyle, 1995). However, there is evidence that more relaxed standards still identify correctly, well fitting models; statistically nonsignificant values for χ^2/df, goodness-of-fit values greater than .90, and RMSEA values below .08 are indicative of models that need not be rejected (Marsh, Hau, & Wen, 2004).

Negative error variance in latent factors can occur and results in inadmissible solutions. One cause is when the sample size relative to the model complexity is small (Chen, Bollen, Paxton, Curran, & Kirby, 2001). If the standard error of the variance exceeds the value of the negative error variance so that there is a strong likelihood that the true value of the error variance includes zero and when the model has been shown to work properly elsewhere it is valid to set the negative error variance to a small positive value (e.g., .005) (Chen et al., 2001). This approach will create an admissible solution and raise the degrees of freedom by one for every error variance so fixed.

Interview. The qualitative results reported in this paper were based on a secondary analysis of the

interview data using categorical analysis (Coffrey & Atkinson, 1996). This style of analysis allows researchers to use both emergent and *a priori* codes to categorize and analyze data. After data were transcribed verbatim, each utterance was labeled using Lankshear and Knobel's (2004) method. For example, in the label L1:032, L referred to the participant pseudonym (Lisa), 1 represented the first interview, and 32 indicated the 32nd utterance within the interview sequence. Initially, the interview data were analyzed phenomenographically (Marton, 1981, 1986) to identify the variation in conceptions present within the data; these results are reported elsewhere (Harris & Brown, in press).

The analysis reported in this paper was specifically concerned with the influence two New Zealand government policy initiatives (i.e., asTTle and AtoL) may or may not have had on teacher conceptions. As such, the coding began with two *a priori* categories: asTTle and AtoL. All passages relating to these categories were cut and pasted into separate documents for further analysis; data on either side of the removed passages were also taken to provide context for the utterances. The data were read multiple times and emergent analytic coding was used to capture key themes.

Within the asTTle data, five subcategories were developed relating to asTTle. The first three were related to asTTle use: (1) new user (those who had undergone training, but had not personally utilized asTTle), (2) reporting (those using asTTle for reporting and/or student tracking or streaming), and (3) improvement (those using asTTle data for diagnostic purposes who reported acting on these data). Subcategories four and five captured teacher evaluative statements about their experiences with asTTle: (4) positive outcomes and (5) negative outcomes. Within the AtoL data, three subcategories centered on perceived changes caused by the program: (1) positive changes (those viewed as positive and sustained), (2) negative changes (those described as unhelpful), and (3) unsustained changes (positive changes that ceased once the program finished). These emergent codes explained the main variations within the data and were used to organize the discussion of results.

Results

Teachers' Conceptions of Assessment

The measurement model for the teachers' conceptions of assessment, using all 161 respondents had acceptable fit characteristics (χ^2 = 632.97, df = 313, χ^2/df = 2.02, p = .16; gamma hat = .87; RMSEA = .08 [90%CI .071-.089]), suggesting that, notwithstanding the small sample size, the survey was able to accurately estimate the complex model. Note that the error variances for two latent factors was negative and were fixed to .005, resulting in increased degrees of freedom in this model relative to previous results. Nonetheless, this provided sufficient warrant to adopt the model and analyze the relative strength of the four major conceptions of assessment.

It is worth noting that all but one of the six factor intercorrelations had values for the MTAP teachers that differed statistically from the national survey of primary teachers (Brown, 2006). The intercorrelation of irrelevance and student accountability differed only by chance, whereas the MTAP teachers had much stronger intercorrelations between (a) student accountability and school accountability, (b) student accountability and improvement, and (c) school accountability and irrelevance. In contrast, the MTAP teachers had much weaker intercorrelations between (a) school accountability and improvement and (b) irrelevance and improvement. This indicated that for the MTAP teachers, the improvement conception was more about student accountability, much less about school accountability, and more about assessment being irrelevant. Furthermore, for the MTAP teachers, school accountability was much more about student accountability and irrelevance.

Statistics for each conception of assessment factor were found according to the TCoA-IIIA factor patterns. The effect size difference (Cohen's *d*) was determined, using group size weighting to evaluate the scale of mean differences. The mean agreement for the four conceptions of the MTAP teachers was contrasted with the mean scores of 111 Auckland primary teachers surveyed in 2001 (Brown, 2002), and the weighted means of nearly 1,000 New Zealand primary and secondary teachers (Brown, 2007) (Table 1). This comparison establishes whether the Auckland MTAP teachers in this study differed from other Auckland and New Zealand teachers.

Table 1
Conceptions of Assessment Statistics between Willing to be Interviewed and Actual Interview and those Surveyed Nationally and in Auckland

	MTAP Participants 2008		Auckland Primary 2001			NZ Primary & Secondary 2001 & 2007		
N	161		111			977		
	M	*SD*	*M*	*SD*	d^1	*M*	*SD*	d^2
Student Accountability	3.86	.75	3.57	.76	.40	3.70	.91	.18
School Accountability	4.50	.67	2.89	.87	2.12	2.69	1.03	1.85
Improvement	3.63	.50	3.78	.54	-.29	4.07	.68	-.67
Irrelevance	3.48	.38	3.01	.66	.92	2.93	.69	.86

Note: *d* = Cohen's standardized effect pooled by group size; positive values of *d* indicate the MTAP group gave more agreement, negative values indicate MTAP group gave less agreement; d^1=difference between MTAP group and Auckland primary teachers; d^2= difference between MTAP group and those national samples of primary and secondary teachers.

The MTAP teachers differed considerably from the New Zealand and Auckland means for two of the conceptions, but not in the ways that were expected. The MTAP teachers agreed substantially more that assessment was about school accountability and that assessment was irrelevant. The MTAP teachers were more like the Auckland sample than they were to the New Zealand sample in terms of the conceptions that assessment was for improvement. This may be indicative of a regional effect in which improvement is less of a clear purpose for assessment within the Auckland metropolis than it is elsewhere in the country; however, this is an unexplored and unexpected result. It is clear that that the MTAP group agreed only slightly more than the other groups with the student accountability conception and this difference is relatively trivial in contrast to the other substantial differences. On the whole, the MTAP teachers predominantly conceived of assessment as a means of school accountability, which itself was not about improvement while being more irrelevant, disproving the hypothesis that the provision of improvement-oriented resources like asTTle and AtoL would lead to higher agreement with the improvement conception.

Given the interviewee selection process, the scale of mean score differences by conception of assessment between those willing to be and actually interviewed was reasonably small (i.e., effect sizes ranged from *d* = .03 to .31) (Brown & Harris, 2008). Hence, it was concluded that the interviewee selection process did not greatly bias the results and that insights from their responses may legitimately shed light on how the introduction of asTTle and AtoL has contributed to these unexpected changes.

Assessment Practices that Define Assessment

Three conceptual factors were found that expressed the types of assessment teachers used when defining the term "assessment." These were classified as oral (i.e., oral question and answer, conferencing, and unplanned observation), tests (i.e., teacher-made written test, standardized test, essay test, and one to three hour examination), and classroom interaction (i.e., planned observation, student written work, group or individual projects, student self- or peer-assessment, and portfolio/scrapbook). The measurement model, consisting of three first-order factors and a second-order factor of assessment practices had sufficient fit to suggest usage for research purposes ($\chi^2 = 177.85$, $df = 52$, $\chi^2/df = 3.42$, $p = .06$; gamma hat = .89; RMSEA = .12 [90%CI .10-.14]). The average selection rate for each of the categories was similar (i.e., oral: $M = .73$, $SD = .37$; test: $M = .76$, $SD = .31$; classroom: $M = .77$, $SD = .32$), suggesting that each type of practice could be considered assessment. However, the regression paths to oral ($\beta = .91$) and classroom ($\beta = .97$) were similar and twice the strength to tests ($\beta = .48$), indicating that tests were not the dominant practices defining assessment. Nonetheless, test-like assessments were not excluded from practices that define assessment.

Structural Model Linking Assessment Conceptions to Definitions

As mean scores alone are insufficient to understand the conceptions of this sample, a structural equation model using teachers' conceptions of assessment (Brown, 2006) and their definitions of assessment was created to examine the interrelationships among factors. The model was designed on the basis that the four major conceptions act as precursor beliefs to how teachers would define assessment. Hence, the four conceptions were regressed simultaneously onto the second-order factor of assessment definitions. The model had acceptable fit to the data considering the sample size was only 161 ($\chi^2 = 1267.36$, $df = 685$; $\chi^2/df = 1.85$, $p = .17$; gamma hat = .84; RMSEA = .073 [90%CI .067-.079]). The pathways from the four conceptions were statistically nonsignificant; thus, reanalysis to find the maximum number of statistically significant predictors of assessment definitions was conducted. Only one pathway could be found that was statistically significant (Figure 2) with marginally better fit ($\chi^2 = 1270.10$, $df = 688$; $\chi^2/df = 1.85$, $p = .17$; gamma hat = .84; RMSEA = .073 [90%CI .066-.079]).

The model showed that the improvement conception of assessment was the only predictor ($\beta = .32$) of the three definitions of assessment and that the correlation between the improvement and school accountability conceptions was not statistically significant ($r = .16$). While the mean score for school accountability was highest, the model made it clear that the school accountability conception was independent of the improvement conception, which itself was the only predictor (explaining some 10 percent of variance) of the practices teachers associate with assessment. The relationship of the three categories of assessment definitions remains unchanged—the improvement conception predicted assessment as defined primarily by oral and classroom practices.

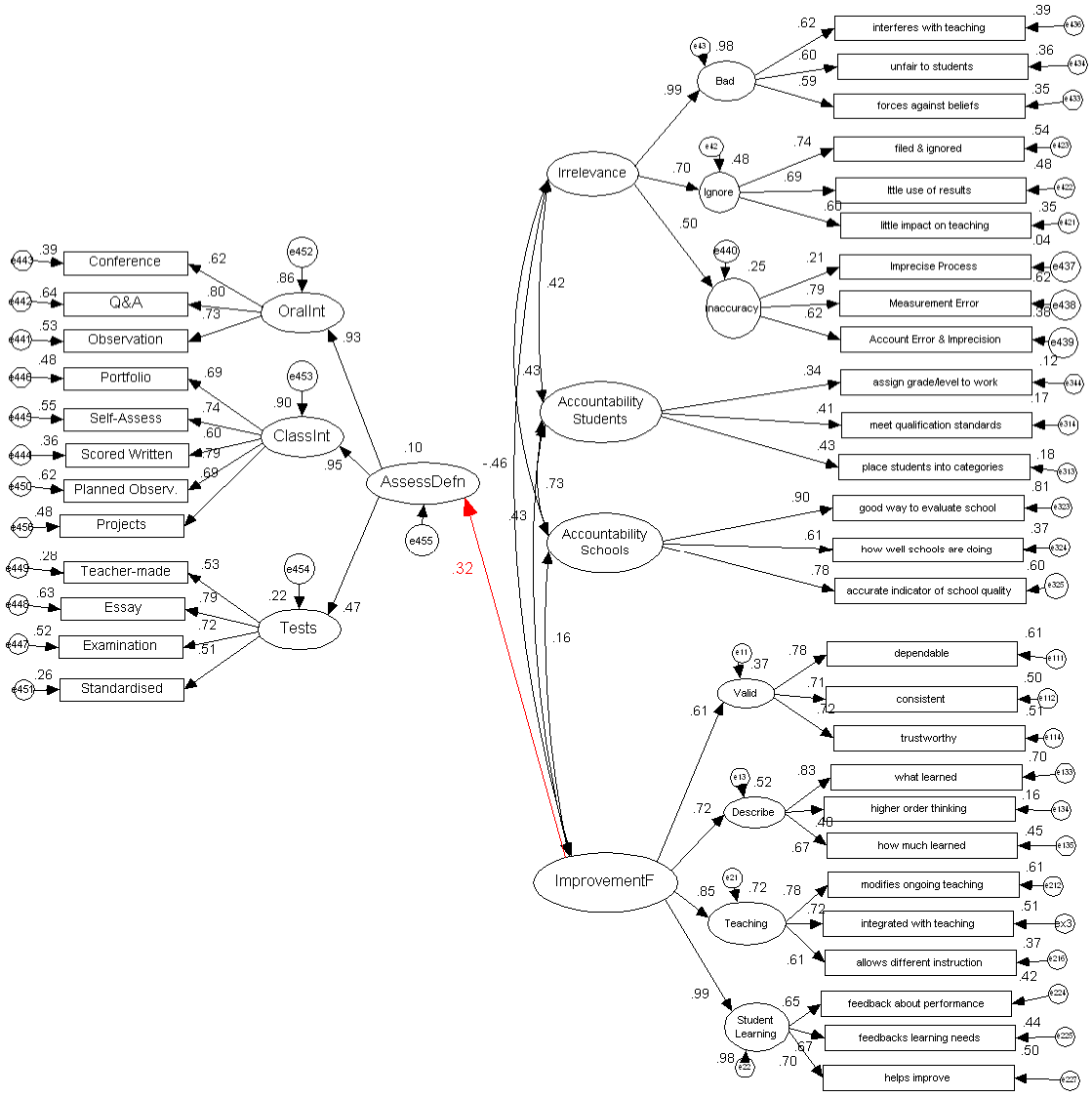

Figure 2. MTAP Study 1 Conceptions of Assessment Related to Definitions of Assessment

Summary of the Survey Results

The survey results appear to be counterintuitive, given that the two policies implemented nationally should have raised improvement conceptions of assessment scores. Instead, the teachers had a high mean score to school accountability conception of assessment which had a nonsignificant relationship to improvement. Furthermore, the improvement conception was the only statistically significant predictor of teachers' assessment definitions that emphasized the use of oral and classroom assessments, without excluding the use of test-like practices for the same purpose. This suggested that the sample did not see improvement and accountability as compatible purposes (even though in previous studies these were correlated). While they defined their personal assessment practices as being primarily oral/interactive (which they saw as related to improvement), their strong agreement that assessment was about school accountability indicated that they also experienced

"assessment" as unrelated to improvement. Perhaps the weaker role test-like practices played vis-a-vis improvement was a reflection of the joint use of tests for improvement and accountability. The interview results were used to determine potential causes of this unexpected outcome in the school-based implementation experiences of the reform initiatives.

Analysis of Interview Data

The goal of the interview analysis was to identify possible explanations in individuals' stories of how asTTle and AtoL were actually implemented and understood in their own schools. It was hoped that these data would shed light on why teachers in this sample had a strong commitment to school accountability, despite experience with policy resources (i.e., asTTle and AtoL) that should have stimulated greater agreement with the improvement conception.

Interviews showed that teachers did not have equal access to these initiatives. While teachers at all schools described some access to asTTle, eight out of twenty-six were classified as "new users," indicating limited personal experience with implementation. Teachers cited technological obstacles (e.g., password/software problems, limited computer access), late adoption of the program (three of the schools only began using asTTle in 2008), or use of other assessment tools (e.g., Progress and Achievement, Star) as reasons for their limited use. Seventeen of the twenty-six participants had not taken part in AtoL, as their schools had not applied for and/or won this contract. Data suggested that demand outstripped supply, as at least one school involved in this study had unsuccessfully applied for the initiative in 2008. Hence, these analyses were based on small samples (asTTle $n = 26$; AtoL $n = 9$).

asTTle. While not all interviewed teachers actively used asTTle, all had been exposed to it. Interview data showed that schools and teachers were utilizing it in diverse ways. The new users ($n = 8$) were generally unclear about exactly how data would be used. For some, asTTle was described as another standardized test to record and report without considering improvement purposes ($n = 5$). Thirteen teachers described utilizing asTTle in ways more aligned with the improvement purpose. They talked about using asTTle data to diagnose where students needed further instruction and plan learning goals. However, these improvement uses did not preclude data from simultaneously being used for accountability purposes, with or without the teacher's approval, leading to some mixed reactions.

Half of the twenty-six teachers described using asTTle for improvement purposes, albeit in differing ways. One way was for personal teacher use in diagnosing student strengths and weaknesses and evaluating how to help them improve. For example, one primary teacher said, "asTTle's great. The reading's great because it creates my groups for me, gives me all my planning, … it gives me the journals, even worksheets, which is great, and it's catering solely for that student, like you know that that's where they should be, so in that aspect it's really good" (I1:144).

Here, she cited using a range of the resources attached to the asTTle program in order to help improve students individually based on diagnosed weaknesses. This use was seen as positive and purely for student improvement.

However, most teachers said their schools also examined asTTle data schoolwide. For example, one intermediate teacher explained:

> So that's what we need to do, obviously something like asTTle … it doesn't necessarily test what you've taught…. It gives that summative, "Okay as a team, these were our objectives, and we've got this group that's made no progress, this group that's made little progress, you know, what's going on here? Was it the way we taught it? Let's go and have a closer look."(U1:118)

Here, asTTle was described as useful for cross-class analysis of whether school objectives were achieved. Her description that asTTle tests don't "necessarily test what you've taught" seems at odds with its capabilities that allow teachers to personally design tests. However, in a school- or syndicatewide context, it is possible that the alignment between teaching and the administered test becomes vaguer. While this teacher said this use of asTTle had an improvement focus, it was also clearly about accountability as well. As another intermediate school teacher explained:

> It's easy in your little box to do your own thing, but now we are looking at enhanced use of asTTle a bit more across the school… [so] we are all working at a similar standard. The literacy team coordinator has had to write a literacy goal for the school … and everyone has to work towards that goal. Part of the goal is to lift your children two sublevels by November 14th. It's dated. (Y1:064)

While this schoolwide goal clearly had an improvement focus, it also could be seen by some as a threat to teacher autonomy because it did not allow for doing "your own thing." While it is unclear what consequences, if any, existed for those classes and students not meeting the goal, being held to this fixed target may cause some to view asTTle assessment negatively.

Other uses also seemed to blend accountability and improvement purposes. For example, one intermediate teacher explained,

> the [asTTle] tests then get used at three way conferences where we talk to the parents about it…. I talk to the kids about … how to interpret the little graph, and then I talk to the parents about that at three-way conferences, and then it gets put online onto Knowledge Net. So like I've done the posttest on stats after I did my conference so the parents, then can go online and actually see whether the kids have improved…. So it's there for the parents to look at as we get the results. It's there for the kids to actually look at 'cause they set goals. They can then go back and reflect and look at why they improved and what new [material] they have learned and if they haven't improved, what are we going to do? (R1:042)

This passage is another example of asTTle data being used simultaneously for accountability and improvement purposes. While scores and results were being reported to parents and students (hence accountability), this teacher described how they were actively being used by teachers, parents, and students for diagnostic purposes and goal setting. However, later within the interview, she did comment that increased parent access to scores and diagnostic information potentially put teachers under scrutiny, prompting her to adopt practices designed to track the instruction and feedback she gave students.

When schools utilized asTTle without a clear emphasis on student improvement, it appeared that teachers viewed it as relating to accountability or irrelevance purposes. For example, one high school teacher explained,

> Yeah, our school is pretty big on asTTle.… I've got all their grades from last year which is where they were placed and they did another lot at the beginning of this year. Teachers get annoyed about it because it's extra work, takes them out of teaching time, and you don't see where it's going, and all that sort of stuff, but the school is really pushing it so, um, what we're doing with it yet, I'm not too sure. (F1:032)

This teacher claimed he and his colleagues lacked clear directives about how the data could or should be used and articulated that this lack of purpose caused him and other teachers to question its utility.

While the data showed obvious tensions between improvement and accountability purposes, some teachers appeared to have resolved these issues. For example, one primary teacher explained,

> asTTle we use all the time… And that I've found [it] really useful for reporting, but also for grouping the kids and trying to be specific

> about, you know, what the strategies are to work on. (O1:088)

Here, the same test is seen as useful for reporting, but also for deciding what needs to be taught next. However, others did not see these two uses as compatible. As one primary teacher explained,

> As a teaching tool, I think it [asTTle] is quite useful. But we discovered that by letting the children have just the limited forty minutes that your slower children are not finishing and asTTle reports that what was not finished was not known, which is not necessarily true. So what we're trying to do is have that because that's a requirement for reporting, so we do that absolutely accurately, but then we give them another test which we allow them to go to the end of and then we use that for teaching because that will tell us what the children don't know. (M1:006)

While she is clearly willing to use test data for improvement purposes, this is only when she has been satisfied it is valid and reliable, something she indicated was not achieved when time limits were used.

However, some teachers seemed to view standardized tests as inherently flawed, making them incompatible with improvement. For example, one primary teacher explained,

> In the e-asTTle, there is potential I believe for them to guess. So they could just go eenie, meenie, miney, moe and stick in a; although the actual analysis of the test breaks it down into where their difficulties are, if they're guessing, it might not be a true indicator of their reading difficulty or ability. (P1:014)

Questions about the validity and reliability of asTTle were raised, in addition to concerns about how data were used for streaming and tracking. With asTTle writing, there was concern cited over the comparability of marks. These are similar concerns to ones teachers have expressed about tests in general (e.g. Harris & Brown, in press), showing that for some teachers, perhaps asTTle will never be seen as an improvement tool because of the firm entrenchment within their thinking that tests are about accountability and lack validity.

One of the strengths of asTTle that was seldom described as utilized was the program's ability to generate personal tests for individual students and classrooms. For example, as one primary teacher explained,

> Well, I like asTTle. Because I've only just finished being trained, I did an assessment paper and we learned how to use the asTTles and what I like about them is that they are more targeted to your school, or your community or your demographic and you can still use those higher order questions, but you focus them into your kids. So I do like the idea, but when we don't make the tests, it's irrelevant to me. It's just a normal standardized test then, so unless they're made by you for your specific needs then they're just another test. (X1:022)

He then went on to explain how because of computer access issues, teachers at his school have been unable to create their own tests, instead having to utilize the premade asTTle tests that he described as irrelevant. Poor computer access and pushes within schools to have students take identical tests so they are "comparable" seem to cause this feature to be used infrequently.

These data suggest that while asTTle was designed with an improvement purpose in mind, the way it is being implemented in schools is often heavily focused around accountability in tandem with or quite apart from improvement purposes. While designed as a low-stakes diagnostic tool, it is clearly being used more for accountability in some syndicates and schools. Additionally, the majority of the features which make it a personalized improvement tool (e.g., the ability for individual teachers to design their own tests, the presence of free resources aligned with diagnostic data) are seldom reported as used. Data suggest that in some schools asTTle use has led to increased accountability and that poor implementation may cause teachers to view the tool as negative

or irrelevant. Additionally, it seems that some teachers still struggle with the concept that testing can be a vehicle for improvement, as this is so different to the student engagement model of assessment *for* learning (e.g. Black & Wiliam, 1998).

Assess to Learn. Assess to Learn was only discussed by nine teachers within the sample of twenty-six, consistent with the restricted provision of this initiative. Of those participating, opinions were mixed. While most gave some positive feedback about the initiative, four mentioned changes disappearing once the initiative ended.

Four teachers were extremely positive in their evaluation of the program. One primary school teacher who referred to herself as "converted" explained,

> What I've learned is that it's very empowering for a child. But … unless that's followed up in the next year, some of that really good stuff can go to waste, and so I think one of the important things is to really get teachers that professional development that they need. It changes their pedagogy and changes the way that they do things. (G1:120)

While this teacher highlighted how valuable she felt the course was in changing teacher pedagogy and actions, she noted that without follow-up, it can "go to waste."

Five of the nine teachers specifically talked about how the program can change teacher thinking or practices. For example, one high school English teacher explained,

> I feel like there's actually quite a considerable shift occurring in teaching pedagogies and the movement away from a teacher directed classroom to co-constructed learning…. I think staff here are at a range of different points along the continuum in both their practice and their understanding, and I think often your practice and your beliefs or your understanding are not necessarily in the same place…. I think those [AtoL] observations and the feedback, I think that is creating a degree of shift…. And to some extent it's occurring in the different classrooms, but … I've only observed three or four other teachers, and three of those are very much co-constructed, you know, and very student-directed and do seem to be working in the way recommended by the AtoL PD. (N1:058, 60)

This teacher evaluated the initiative positively, stating it was at least partially responsible for a shift towards an improvement orientation towards assessment. However, she also raised two important points relating to the implementation. First, she noted that teachers' practices and beliefs were not necessarily congruent. Second, she mentioned that while she had observed three teachers who did seem to be fully adopting AtoL in their classroom practice, she didn't mention the fourth teacher, implying that this person may not have taken the strategies on board.

These data suggest that AtoL is being taken up differently by teachers, a point confirmed by another high school teacher at the same school:

> We have this assessment [for] learning person come from the ministry who comes and observes our classes…. she wants to see things like success criteria written up on the board and learning criteria and stuff like that…. I've been an educator for quite a few years now, and I've seen a lot of ideas pushed that disappear into nowhere again, and basically everyone just carries on and teaches in their teaching style that they're really good at. (Q1:218)

These data suggest that this teacher had no intention of changing his practices to align with AtoL as he likened it to other programs that disappear with time, leaving teachers to revert back to their preferred style. This particular teacher seemed to view these reforms as irrelevant.

The most frequently cited problem with the AtoL course was that many positive changes disappear once the initiative finishes. As one intermediate teacher noted,

> I was lucky enough as I came into the school, I was put into the assessment team and Assess to Learn. I did a course on Assess to Learn, but

> somehow [after the course finished] that team just fragmented and went, it [was] lost. We don't have an assessment team anymore in the school running as such, like when I came. So for some reasons, I think people were not onto it. (D1:174)

Another primary teacher cited that while it did introduce some good ideas into the school, it was not as effective as it could have been:

> It probably was one of those things that was conceived in all of the best spirit of the thing, but was then dumped within a school that has realities…. It's one of those kinds of catch-all things that we were quite pleased to see the back of. And as I said, a lot of good came out of it, but there was a lot of stuff. We rewrote our school's scheme entirely around [it], which was a load of rubbish. Got rid of it again. (M1:062, 64)

While this teacher acknowledged that some good came of the initiative, she indicated that the majority of staff were quite happy when it finished and suggested that on completion many things reverted to as they were before.

Overall, these data provide some insight into the effect that AtoL may be having on teacher thinking. First, these data highlight that only a small number of teachers actually have access to this training, making it questionable how much of a global impact it could have on the thinking of this sample of teachers. While five teachers cited it could change teacher thinking and practices in positive ways, four identified that changes were not sustained. These data also show that some teachers choose to resist the changes, seeing AtoL as another initiative that will disappear over time; for these teachers it is possible that this initiative might encourage them to adopt an irrelevance conception towards assessment. Hence, while this initiative was designed to encourage teachers to adopt an improvement conception, lack of access, teacher resistance, and lack of follow-up may explain why it has not led to heighten conceptions of assessment as improvement.

Discussion

Together these data show a group of teachers struggling with the combined tensions of assessment for improvement and assessment for accountability. High mean scores for the school accountability conception appeared to be driven by two groups of teachers—(a) those who saw assessment as primarily a negative means of demonstrating school competence and quality and (b) those who saw assessment as a legitimate means of improving instruction and demonstrating accountability. It would also appear that these two groups are quite independent of each other (consider the low correlation between the two conceptions). This raises the possibility that effective provision of improvement resources (e.g., asTTle and AtoL) enables some teachers to use assessment for both improvement and accountability. In contrast, more teachers appeared to associate testing with a negative external accountability process.

From the interview data, it is possible to identify a number of potentially contributing factors. First was the issue of access. The use of asTTle, as per policy, is voluntary; this means that not every teacher within a school using asTTle would necessarily utilize it. Furthermore, there are no external regulations as to how asTTle should be used; this is determined by school policy. That policy can be to use asTTle primarily for schoolwide accountability purposes or for individual class improvement goals; the choice is with the schools and data show a variety of positive and negative ways in which it appears to be being implemented. The interview data showed that only about a third of the teachers had been involved in the AtoL program; not every school that wanted the assistance was able to obtain it. Thus, lack of support and a high degree of flexibility in New Zealand school governance may have contributed to the unexpected results found in this survey.

A second issue was the wide range of reactions to these tools. With asTTle, while some teachers appeared to be able to reconcile the use of the same tool for accountability and improvement purposes, others could not. Additionally, perhaps due to the push for interactive assessment in recent years, some struggled to believe that test results could ever be valid and reliable and therefore truly useful for improvement. With the AtoL program, while some teachers were clearly committed to changing their practice, others viewed it as a passing fad which could safely be ignored. It remains unclear what fundamental differences in these teachers could cause these diverse reactions. In particular, it would be useful to examine more extensively the thinking of those who do seem to be able to see improvement and accountability as being aligned.

The number of teachers participating in AtoL was relatively small and unlikely to contribute significantly to the shift observed through the questionnaire data. However, all participants cited some use of asTTle; asTTle school-based implementation was associated highly with accountability, in many cases more so than improvement. Teacher's reactions to this school-determined emphasis may help to explain the strong agreement that assessment was about school accountability. Likewise, the poor implementation in some schools could have contributed to the elevated agreement with the irrelevance construct. However, due to the small sample size of interviewees, these explanations are tentative. What these interview data clearly illuminate, however, are that schools and individuals mediate the implementation of any policy initiative and therefore cause it to have a range of often unintended consequences.

Another interesting finding was that only the improvement conception predicted teachers' definitions of assessment. Teachers in this study defined assessment as primarily oral and interactive practices. Testing was still included in assessment definitions, although not as strongly, suggesting tests are viewed as problematic for improvement. Tests can, in the views of the MTAP teachers, be used for improvement (see examples above of how asTTle can achieve this), but they tend to be predominantly viewed as student and school accountability measures. The interview results suggest that this conception was not a prejudice on the part of all teachers, but at times related to how the asTTle test system was actually being implemented. Instead of being primarily an improvement system with accountability information, the test-like nature of the resource was being used primarily for school quality or accountability functions and this practice influenced the views of teachers in the schools. At best, the formative intentions of using the asTTle test system to improve teaching and learning were partially fulfilled, especially among teachers who experienced the effect of asTTle on their own classroom teaching. The perception that tests are inherently accountability measures is not dominant; tests can support diagnostic, formative evaluations, but it would appear the policy priorities of schools can subvert such usage.

That only the improvement conception predicted the various definitions of assessment suggested that perhaps some teachers adopted an implicit theory of assessment in which *good schools use assessment to improve provided assessment is defined as predominantly informal and interactive practices*; although, a few teachers would allow test-like practices into the mix. This suggests that there may be difficulties in getting past this conception if a policy adopts tests as means of supporting assessment *for* learning. This result is in alignment with Brown (2009) in which it was found that school accountability conceptions positively predicted the use of cognitively deep assessments, whereas test-like practices and surface cognitive demand were predicted by the student accountability conception. That same study found that both irrelevance and improvement conceptions predicted the use of informal, classroom assessment practices. Thus, there appears to be some consistency in

teachers' reluctance to view tests as a legitimate assessment *for* learning practice, even when testing systems are explicitly designed to do so. The current study suggests that school accountability has become dominant in teacher thinking and that the current provision of assessment improvement resources and professional development has not yet persuaded the majority of these teachers that assessment can meet both improvement and accountability requirements. Unlike previous studies with the TCoA in New Zealand and Australia (Queensland), the school accountability conception was positively correlated, not with improvement, but rather with irrelevance. This result should give pause to advocates of policies that prioritize assessment *for* learning through external or standardized tests.

This study shows that there are considerable obstacles to overcome in the minds of teachers related to the implementation of policies and resources related to assessment for learning. While professional development may appear a promising solution for promoting assessment *for* learning, teacher responses to AtoL from this study suggest that many changes dissipate once the program finishes. The importance of schoolwide adoption under the direction and support of the school's instructional leadership cannot be understated (Robinson, 2007; Timperley, Wilson, Barrar, & Fung, 2007). Additionally, professional development of this kind has a high fiscal cost, making it unlikely that it will be delivered to all schools and teachers. Alternatively, using tests to improve learning, while relatively economic (Linn, 2000), is unlikely to be seen by many teachers as an obvious component of improved teaching and learning. Although, asTTle overcomes such reservations through its multifaceted, individualized reporting solutions, the power of the system may be overwhelmed by a strong emphasis on assessment for school improvement and the under-utilization of its student and classroom features. While school improvement through analysis of assessment data is important, the public nature of school improvement appears not to have the same conceptual meaning to teachers as helping individual students better their learning. Though there should be synergy to the two agendas (i.e., a good school improves individual students' learning), it would appear that most teachers are yet to be persuaded that both agendas mean the same thing.

While data from this study cannot identify more optimal tools or programs for promoting assessment for learning, they certainly highlight that teacher thinking must be taken into account when enacting policy changes. These data suggest that before any reform is undertaken, teachers must first be persuaded of its utility and allowed to "buy in" to the program. Without this teacher support, reforms or resources are unlikely to maximally achieve their intended goals. Hence, policymakers must work to identify the pre-existing beliefs held by teachers and, when possible, align new reforms and resources to take advantage of the positive outlook teachers have. It is clear from this study that teachers want to use assessment for improved teaching and learning, and this must not be forgotten by policymakers. When reforms cannot, for whatever reasons, be aligned with teacher thinking, extensive efforts will be required to convince teachers of the benefits otherwise implementation is likely to fail.

It is possible that the results of this study are an artifact of the sampling, history, or mortality. Nonetheless, the results are suggestive of the important processes that need to be considered when implementing a complex policy like assessment for learning. While test systems like asTTle can certainly be used for student improvement, this study suggests that schools and teachers may often assume test data are primarily for school accountability. This emphasis, while legitimate, tends to overwhelm the much more fragile conception that assessment is for improved learning. This study draws attention to the tension between the

classroom improvement and schoolwide accountability purposes. If school leaders can be encouraged to publicly prioritize the classroom improvement purpose and if teachers can be persuaded that tests do contribute to improvement rather than just accountability, it is likely that improvement-oriented resources like asTTle and AtoL will achieve assessment *for* learning. Unfortunately, the current study suggests that schools may be rejecting, ignoring, or thwarting the improvement-oriented ambitions and capabilities of these tools. The remaining issue is how to introduce a reform policy in such a way that its implementation enhances and maintains a robust improvement-oriented conception of assessment.

References

Ajzen, I. (2005). *Attitudes, personality and behavior* (2nd ed.). New York: Open University Press.

Asch, R. L. (1976). Teaching beliefs and evaluation. *Art Education, 29*(6), 18-22.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7-74.

Black, P., & Wiliam, D. (2004). The formative purpose: Assessment must first promote learning. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability* (pp. 20-50). Chicago: NSSE & University of Chicago Press.

Brown, G. T. L. (2002). *Teachers' Conceptions of Assessment.* Unpublished doctoral dissertation, University of Auckland, Auckland, NZ.

Brown, G. T. L. (2004a). Measuring attitude with positively packed self-report ratings: Comparison of agreement and frequency scales. *Psychological Reports, 94*, 1015-1024.

Brown, G. T. L. (2004b). Teachers' conceptions of assessment: Implications for policy and professional development. *Assessment in Education: Policy, Principles and Practice, 11*(3), 305-322.

Brown, G. T. L. (2006). Teachers' conceptions of assessment: Validation of an abridged instrument. *Psychological Reports, 99*, 166-170.

Brown, G. T. L. (2007, December). *Teachers' conceptions of assessment: Comparing measurement models for primary and secondary teachers in New Zealand.* Paper presented at the annual conference of the New Zealand Association for Research in Education, Christchurch, NZ.

Brown, G. T. L. (2008). *Conceptions of assessment: understanding what assessment means to teachers and students*. New York: Nova Science Publishers.

Brown, G. T. L. (2009). Teachers' self-reported assessment practices and conceptions: Using structural equation modelling to examine measurement and structural models. In T. Teo & M. S. Khine (Eds.), *Structural equation modelling in educational research: Concepts and applications* (pp. 243-266). Rotterdam, NL: Sense Publishers.

Brown, G. T. L., & Harris, L. (2008, November). *Teachers' conceptions of assessment: Interview validation of survey responses*. Paper presented at the annual conference of the New Zealand Association for Research in Education, Palmerston North, NZ.

Brown, G. T. L., Irving, S. E., & Keegan, P. J. (2008). *An introduction to educational assessment, measurement, and evaluation: Improving the quality of teacher-based assessment* (2nd ed.). Auckland, NZ: Pearson Education.

Brown, G. T. L., Kennedy, K. J., Fok, P. K., Chan, J. K. S., & Yu, W. M. (in press). Assessment for improvement: Understanding Hong Kong teachers' conceptions and practices of assessment. *Assessment in Education: Policy, Principles and Practice*.

Brown, G. T. L., & Lake, R. (2006, November). *Queensland teachers' conceptions of teaching, learning, curriculum and assessment: Comparisons with New Zealand teachers*. Paper presented at the annual conference of the Australian

Association for Research in Education, Adelaide, Australia.

Calderhead, J. (1996). Teachers: Beliefs and knowledge. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 709-725). New York: Simon & Schuster Macmillan.

Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., & Kirby, J. B. (2001). Improper solutions in structural equation models: Causes, consequences, and strategies. *Sociological Methods & Research, 29*(4), 468-508.

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

Cizek, G. J., Fitzgerald, S., Shawn, M., & Rachor, R. E. (1995). Teachers' assessment practices: Preparation, isolation and the kitchen sink. *Educational Assessment, 3*, 159-179.

Clark, C., & Peterson, P. (1986). Teachers' thought processes. In M. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed.) (pp. 255-296). New York: MacMillan.

Coffey, A., & Atkinson, P. (1996). *Making sense of qualitative data: Complementary research strategies*. London: Sage.

Crooks, T. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58*(4), 438-481.

Crooks, T. J. (2002). Educational assessment in New Zealand schools. *Assessment in Education: Principles Policy & Practice, 9*(2), 237-253.

Darling-Hammond, L. (2003, February). Standards and assessments: Where we are and what we need. *Teachers College Record.* Retrieved August 2, 2005, from http://www.tcrecord.org

Education Gazette. (2002). Better assessment likely. *Education Gazette, 81*(6). Retrieved June 6, 2009, from http:// www.edgazette.govt.nz/Articles/Article.aspx?ArticleId=6155

Guthrie, J. T. (2002). Preparing students for high-stakes test taking in reading. In A. E. Farstrup & S. J. Samuels (Eds.), *What research has to say about reading instruction* (pp. 370-391). Newark, DE: International Reading Association.

Hamilton, L. (2003). Assessment as a policy tool. *Review of Research in Education, 27*, 25-68.

Hamilton, L. S., Stecher, B. M., Marsh, J. A., McCombs, J. S., Robyn, A., Russell, J. L., et al. (2007). *Standards-based accountability under no Child Left Behind: Experiences of teachers and administrators in three states*. Santa Monica, CA: RAND Education.

Harris, L. R., & Brown, G. T. L. (in press). The complexity of teachers' conceptions of assessment: Tensions between the needs of schools and students. *Assessment in Education: Principles, Policy, & Practice*.

Hattie, J. A. C., & Brown, G. T. L. (2004, September). *Cognitive processes in asTTle: The SOLO Taxonomy*. asTTle Tech. Rep. #43, University of Auckland/Ministry of Education.

Hattie, J. A. C., & Brown, G. T. L. (2008). Technology for school-based assessment and assessment for learning: Development principles from New Zealand. *Journal of Educational Technology Systems, 36*(2), 189-201.

Hattie, J. A. C., Brown, G. T. L., & Keegan, P. J. (2003). A national teacher-managed, curriculum-based assessment system: Assessment Tools for Teaching & Learning asTTle). *International Journal of Learning, 10,* 771-778.

Hattie, J. A. C., Brown, G. T. L., Keegan, P. J., MacKay, A. J., Irving, S. E., Cutforth, S., et al. (2004). *Assessment Tools for Teaching and Learning asTTle) manual* (Version 4, 2005 ed.). Wellington, NZ: University of Auckland / Ministry of Education/Learning Media.

Hattie, J. A., Brown, G. T. L., Ward, L., Irving, S. E., & Keegan, P. J. (2006). Formative evaluation of an educational assessment technology innovation: Developers' insights into Assessment Tools for Teaching and

Learning (asTTle). *Journal of MultiDisciplinary Evaluation, 5*(3), 1-54.

Heaton, J. B. (1975). *Writing English language tests*. London: Longman.

Hershberg, T. (2002). Comment. In D. Ravitch (Ed.), *Brookings papers on education policy: 2002* (pp. 324-333). Washington, DC: Brookings Institution Press.

Hoyle, R. H. (1995). The structural equation modeling approach: Basic concepts and fundamental issues. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 1-15). Thousand Oaks, CA: Sage.

Kahn, E. A. (2000). A case study of assessment in a grade 10 English course. *Journal of Educational Research, 93*, 276-286.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to Western thought*. New York: Basic Books.

Lankshear, C., & Knobel, M. (2004). *A handbook for teacher research*. Berkshire: Open University Press.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4-16.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cut-off values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11*(3), 320-341.

Marton, F. (1981). Phenomenography: Describing conceptions of the world around us. *Instructional Science, 10*, 177-200.

Marton, F. (1986). Phenomenography: A research approach to investigating different understandings of reality. *Journal of Thought, 21*(3), 28-49.

McMillan, J. H. (2001). Secondary teachers' classroom assessment and grading practices. *Educational Measurement: Issues and Practice, 20*(1), 20-32.

McMillan, J. H., Myran, S., & Workman, D. (2002). Elementary teachers' classroom assessment and grading practices. *The Journal of Educational Research, 95*(4), 203-213.

Ministry of Education. (1994). Assessment: Policy to Practice. Wellington, NZ: Learning Media.

Ministry of Education. (2001). Developing teachers' assessment literacy. *Curriculum Update, 47*.

Ministry of Education. (2007). *The New Zealand Curriculum for English-medium teaching and learning in years 1-13*. Wellington, NZ: Learning Media.

Monfils, L. F., Firestone, W. A., Hickes, J. E., Martinez, M. C., Schorr, R. Y., & Camilli, G. (2004). Teaching to the test. In W. A. Firestone, R. Y. Schorr & L. F. Monfils (Eds.), *The ambiguity of teaching to the test: Standards, assessment, and educational reform* (pp. 37-61). Mahwah, NJ: LEA.

Noble, A. J., & Smith, M. L. (1994). *Old and new beliefs about measurement-driven reform: "The more things change, the more they stay the same"* (CSE Technical Report No. 373). Los Angeles,: University of California, Los Angeles, CRESST.

Pajares, M. F. (1992). Teachers' beliefs and educational research: Cleaning up a messy construct. *Review of Educational Research, 62*, 307-332.

Peddie, R. (2000). *Evaluation of the assessment for better learning professional development programmes: Final report*. Auckland, NZ: Auckland UniServices Limited.

Richardson, V., & Placier, P. (2001). Teacher change. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed.) (pp. 905-947). Washington, DC: AERA.

Popham, W. J. (2000). *Modern educational measurement: Practical guidelines for educational leaders* (6th ed.). Boston: Allyn & Bacon.

Poskitt, J., & Taylor, K. (2008, July). *National education findings of Assess to Learn (AtoL) report*. Wellington, NZ: Ministry of Education.

Robinson, V. M. J. (2007). *School leadership and student outcomes: Identifying what works and why*

(ACEL Monograph No. 41). Melbourne, AU: Australian Council for Educational Leaders.

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. Harlow, UK: Pearson Education.

Smith, M. L., & Fey, P. (2000). Validity and accountability in high-stakes testing. *Journal of Teacher Education, 51*(5), 334-344.

Stecher, B. M., & Barron, S. (2001). Unintended consequences of test-based accountability when testing in "milepost" grades. *Educational Assessment, 7*(4), 259-281.

Thompson, A. G. (1992). Teachers' beliefs and conceptions: A synthesis of the research. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 127-146). New York: MacMillan.

Timperley, H., Wilson, A., Barrar, H., & Fung, I. (2007). *Teacher professional learning and development.* Best Evidence Synthesis Iteration (BES) Report. Wellington, NZ: Ministry of Education.

Tittle, C. K. (1994). Toward an educational psychology of assessment for teaching and learning: Theories, contexts, and validation arguments. *Educational Psychologist, 29*, 149-162.

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

Warren, E., & Nisbet, S. (1999). The relationship between the purported use of assessment techniques and beliefs about the uses of assessment. In J. M. Truran & K. M. Truran (Eds.), *22nd annual conference of the Mathematics Education and Research Group of Australasia,* Vol. 22 (pp. 515-521). Adelaide, SA: MERGA.

Webb, N. L. (1992). Assessment of students' knowledge of mathematics: Steps toward a theory. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 661-683). New York: Macmillan.