QUEENSLAND TEACHERS' CONCEPTIONS OF ASSESSMENT: THE IMPACT OF POLICY PRIORITIES ON TEACHER ATTITUDES

Gavin T L Brown*

The Hong Kong Institute of Education

Robert Lake

Novum AVI

Gabrielle Matters

Australian Council for Educational Research

Running Head: Queensland Teachers Assessment

*Corresponding Author:

Dr Gavin T L Brown, Faculty of Education Studies, Hong Kong Institute of Education, 10 Lo Ping Road, New Territories, Hong Kong SAR.

Tel: + 852-2948 8529

Email: gtlbrown@ ied.edu.hk

Fax: + 852-2948-7563

Keywords:

Teacher Attitudes; Attitude Measures; Assessment; Educational Policy; Queensland;

Abstract

The conceptions Queensland teachers have about assessment purposes were surveyed in 2003 with an abridged version of the Teacher Conceptions of Assessment Inventory. Multi-group analysis found that a model with four factors, similar in structure to New Zealand studies, was statistically different between Queensland primary and (lower) secondary teachers. Primary teachers agreed more than secondary teachers that 'assessment improves teaching and learning', while the latter agreed more that it 'makes students accountable'. The inter-correlation of 'assessment is irrelevant' to 'makes students accountable' was statistically stronger for primary teachers. Teacher beliefs reflected the differing practices of assessment by level of schooling.

1. Introduction

In this study we use the term 'conception' to capture all that a teacher thinks about the nature and purpose of an educational process and practice (Thompson, 1992). Evidence exists that teachers' conceptions of various aspects of the education process (e.g., teaching, learning, and curricula) strongly influence how they teach and what students learn or achieve (Clark & Peterson, 1986; Thompson, 1992; Calderhead, 1996). Specifically, teachers' beliefs about students, learning, teaching, and subjects influence assessment techniques and practices (Asch, 1976; Cizek, Fitzgerald, Shawn, & Rachor, 1995; Kahn, 2000; Tittle, 1994). This is consistent with Ajzen's (2005) model of planned or reasoned behaviour, which suggests that teachers' intentions, beliefs about what others think, and sense of power to fulfill their intentions determine their behavior within school environments. Teacher belief systems appear to arise from their early experiences of educational processes (Pajares, 1992).

Needless to say, teachers do not exist in a vacuum; they work in social environments and carry out multi-faceted work (e.g., planning, teaching, and evaluating) within a jurisdiction that directs policy concerning curriculum, pedagogy, assessment, and so on. Furthermore, the policy choices within the jurisdiction express the societal and cultural norms valued by members of that jurisdiction, most of whom are not teachers. However, the introduction of a policy reform around assessment (e.g., Australia's *National Assessment Program–Literacy and Numeracy* which tests all students in Years 3, 5, 7, and 9) may express values not necessarily held by those employed to implement the policy (i.e., teachers). Hence, when considering educational reform we should pay attention to the conceptions that teachers have about their current practices in order to appreciate how they are most likely to understand, respond to, and implement the reform.

Teachers' conceptions about their practices may well influence their practices, and their response to reforms that seek to change practice. Although much research has been done on teachers' assessment and grading practices (e.g., McMillan, 2001a; McMillan, Myran, & Workman, 2002; Stecher & Barron, 2001), those investigations have not focused on the purposes or intentions teachers have for their practices. It may be that the various practices described are intended to improve student learning, or it may be that the same practices are carried out in order to fulfil administrative or accountability goals. For example, in a study of New Zealand primary school teachers using items from McMillan's (2001a) questionnaire, it was shown that two different conceptions of assessment (i.e., 'assessment improves learning' and 'assessment is irrelevant') equally predicted the use of interactive–informal assessment practices (Author1, 2009). It is worthwhile, therefore, exploring whether teachers' conceptions are different in different contexts, and whether those putative differences align with characteristics of those contexts or reforms.

As well as a relationship between concept and practice, we can postulate a relationship between context and both the policies that are proposed and the conceptions about those policies. We expect that differences in culture or society lead not only to differences in policy but also to differences in conceptions of corresponding practices and processes: the greater the difference in policy, the more unlike would be the conceptions. For example, Hamilton et al. (2007) reported that teachers in California, Georgia, and Pennsylvania had very similar responses to, experiences in, and attitudes towards standards-based accountability assessments, perhaps because there were very small differences in how the systems were implemented. In contrast, teachers in New Zealand and Hong Kong, which have very different assessment regimes—New Zealand has no public examinations before senior secondary school, whereas Hong Kong uses public examinations extensively (Choi, 1999), had very different conceptions of how grading students was related to the purpose of improvement as indicated by the responses to a common research instrument (Author1, Kennedy, Fok, Chan, & Yu, 2009). In Hong Kong, the notion of assessment as evaluating

students was strongly correlated (r=.91) with the notion of assessment as leading to improvement. In New Zealand, the same two conceptions were weakly correlated (r=.21). The difference was attributed to cultural features of the Confucian system in Hong Kong, which, inter alia, emphasizes educational testing as a force for improving learning.

The model we use (Author1 & Harris, 2009) posits that societal and cultural contexts determine the nature of policy priorities; that policy priorities – including reforms or innovations – affect teachers' beliefs and attitudes; and that these in turn influence the translation of policy into practice in the classroom and school. Student achievement outcomes are a function of policy reforms as implemented by teachers. Although this conceptual framework is very similar to that advanced by Hamilton, et al. (2007), there are two important distinctions. One is that teachers' beliefs are not positioned external to the implementation processes but rather as mediating between policy and outcomes. Another is that policy directions are seen as a function of priorities within society and culture, suggesting that variation in practices and outcomes might be a function of one or more of three factors – teachers, policies, and societies.

1.1. Conceptions of assessment

We suggest that four conceptions of assessment exist, three of which may loosely be categorised as 'purposes' and one as an 'anti-purpose' (Author1, 2008). Three major purposes for assessment thread the scholarly literature (e.g., Heaton, 1975; Torrance & Pryor, 1998; Warren & Nisbet, 1999; Webb, 1992):

- assessment as improvement of teaching and learning (*Improvement*) ;
- assessment as making schools and teachers accountable for their effectiveness (*School Accountability*); and
- assessment as making students accountable for their learning (*Student Accountability*).

What we term an anti-purpose is a belief that assessment is fundamentally irrelevant to the life and work of teachers and students (*Irrelevant*, Shohamy, 2001). The italicised terms are those we shall use in this paper to refer to these four conceptions.

Improvement, sometimes known as assessment for learning or formative assessment, has been shown to have a positive impact on educational outcomes (Black & Wiliam, 1998; Crooks, 1988; Popham, 2000). Formative evaluations, using a wide range of assessment techniques and strategies (Linn & Gronlund, 2000), are carried out by teachers and schools during the process of instruction for the purpose of improving student learning outcomes and teacher instructional practices (Scriven, 1991). It is imperative that the assessments are aligned with curriculum, course planning, and teaching (Firestone, Mayrowetz, & Fairman, 1998; Gipps, et al., 1995), that the assessments are accurate (Thorndike, 1997), as well as valid (Messick, 1989), so that appropriate feedback to the learner and instructor in response to assessed performance is provided (Hattie, 2009; Sadler, 1989).

The *School Accountability* conception uses assessment results to publicly demonstrate that teachers or schools are doing a good job (Butterfield, Williams, & Marr, 1999; Mehrens & Lehmann, 1984: Smith, Heinecke, & Noble, 1999) and imposes consequences for schools or teachers for reaching or not reaching required standards (Firestone, Mayrowetz, & Fairman, 1998; Guthrie, 2002). Two rationales exist: publicly demonstrating that schools and teachers deliver quality instruction; and improving the quality of instruction. The first viewpoint insists that schools and teachers have to be able to demonstrate that they are delivering the quality product that society is entitled to by virtue of funding the educational process (Crooks, 1990; Gipps, et al., 1995; Hershberg, 2002; Smith and Fey, 2000), while the second viewpoint emphasises the role testing can play in improving teacher and student work

(Linn, 2000; Noble & Smith, 1994; Porter & Chester, 2002). This was expressed most succinctly in the *What You Test is What You Get* aphorism (Resnick & Resnick, 1992) who enjoined the goal of building assessments that directed educators to the kind of teaching needed for real improvement in learning (i.e., assessments of higher order abilities).

The central tenet of *Student Accountability* is that assessment holds students individually accountable for their learning through giving of grades or scores, checking off performance against criteria, and reporting grades to parents, future employers, and other educators (McMillan, 2001b). These assessments have high-stakes consequences for students: for example, tracking students into ability streams, graduation, entry selection to higher levels of education, retention in grade, assignment to remedial education tracks, and so on. The act of making public student performance through certification is generally considered legitimate and important (Guthrie, 2002). Some teachers believe that this type of assessment places a necessary and motivating pressure on students (Kahn, 2000); whereas, others believe high-stakes testing has an adverse emotional impact on young students, causing unwarranted worry and anxiety (Guthrie, 2002; Philipp, Flores, Sowder, & Schappelle, 1994), and some suggest that student accountability testing is biased against certain population groups, specifically, low socio-economic and ethnic minority populations (Neill, 1997). Hence, the process of grading or evaluating students generates a mixture of responses, including rejection of assessment.

The conception *Irrelevant* is based on the view that external evaluation processes are inadequate, inaccurate, and/or irrelevant to the teachers' ability to improve student learning. Indeed, while some positive educational effects of external accountability tests are becoming evident (Au, 2007; Black & Wiliam, 2004; Cizek, 2001; Monfils, et al., 2004), most evidence is that high-stakes school accountability measures have negative consequences on curriculum, teachers, teaching, and student learning and that teachers are aware of this (Au, 2007; Darling-Hammond, 2003; Firestone, Mayrowetz, & Fairman, 1998; Hamilton, 2003). Indeed, evidence for the irrelevance of high-stakes testing systems to valued educational outcomes can be seen in wide-spread test score inflation in the United States (Koretz, 2002; Linn, 2000), suggesting that assessment is not working to improve or demonstrate quality. There is even a view, most evident in England and other Commonwealth nations, that teachers' intuitive, intimate, and continuing knowledge of student learning is the best basis for improving and reporting on student learning and, consequently, formal assessments are not relevant or needed (Asch, 1976; Cooper & Davies, 1993; Gipps, et al., 1995; Torrance & Pryor, 1998).

One of the difficulties in researching teachers' conceptions of assessment is that they appear to hold multiple and possibly even contradictory conceptions without being disturbed by such contradiction (Cizek, Fitzgerald, Shawn, & Rachor, 1995; Kahn, 2000). Human belief systems appear capable of storing separate, contradictory representations of an object or process (see discussion in Author1, 2008). A major factor in this plurality of conceptions is that assessment itself serves multiple purposes, which may be complementary or contradictory. For example, believing that assessment is for improved teaching and learning (*Improvement*) is commonly understood as being opposed to accountability or evaluation purposes, but this may not be the case if teachers accept the legitimacy of accountability mechanisms. Analysing the interaction of teachers' responses to four conceptions of assessment rather than treating them as automatically mutually exclusive opposites permits a richer, less simplistic, representation of the complexities in how assessment is understood and evaluated. It may also identify how differences in context and policy influence teachers' conceptions of assessment.

Author1 (2008) showed that, among a sample of over 500 primary school teachers in New Zealand, *Irrelevant* correlated with *Student Accountability* (r=.36, $p<.01$) but not with

School Accountability (r=-.13, p>.01), suggesting that, within the low-stakes, child-centred discourse of primary schooling, teachers considered grading students to be bad, unfair, or inaccurate. Furthermore, in a sub-sample of such teachers, Author1 (2009) reported that there was a low level for *School Accountability* (M=2.63, max=6) but it was correlated (r=.50, p<.01) with *Improvement*. At the same time, *School Accountability* was only weakly associated (β=.18) with the use of measures of deep cognitive processing. This suggested that teachers believed schools could be judged if deep student learning was assessed, while assessments of surface processing would not be legitimate measures of school quality. Thus, the advantage of conceptualising teacher beliefs about assessment in terms of their simultaneous attitudes towards four contrasting and/or complementary purposes of assessment is that we can develop a profile of their joint conceptions rather than simply a dichotomous portrayal of teacher opinions as being either *for* one purpose of assessment and automatically being *against* a seemingly contradictory purpose.

1.2. Policy effects on conceptions of assessment

We are then left to consider how these four conceptions might be related within differing policy contexts. We expect that a policy that prioritizes assessment's educational improvement purpose, as is the case in New Zealand (Ministry of Education, 1994), would co-exist with teachers who are strongly committed to *Improvement* conception (Author1, 2004a). 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 a narrowed range of student learning outcomes to inappropriate external testing (Darling-Hammond, 2003; Hamilton, 2003; Hamilton et al., 2007; Linn, 2000). Recent reports, however, provide evidence of the positive consequences of national testing (e.g., Black & Wiliam, 2004; Cizek, 2001; Monfils et al., 2004) and also of teachers being aware of the link between accountability pressures and educational improvement (Hamilton et al., 2007). It therefore seems that, in the minds of teachers, the conceptions of *School Accountability* and *Improvement* are not simple opposites.

Research on New Zealand teachers shows that within the context of a policy framework emphasizing assessment for improvement and entrusting schools with the responsibilities for monitoring and reporting progress to parents and government, teachers emphasize the two conceptions of assessment – *Improvement* and *Student Accountability* (Author1, 2008). However, the correlations between each of *Student Accountability* and *School Accountability*, and *Improvement* were intriguing. Whereas teachers had high means for *Improvement* and low means for *School Accountability*, there was a moderate positive correlation between *Improvement* and *School Accountability* (r=.46). In contrast, there was a similar moderate correlation between *Student Accountability* and *Irrelevant* (r=.36). Furthermore, *School Accountability* predicted the use of deep learning assessment practices; whereas, *Student Accountability* predicted the use of surface learning and test-like assessment practices (Author1, 2009). Thus, *Improvement* is prioritized in the low-stakes assessment policy framework of New Zealand. Taken together, these patterns suggested that teachers wanted to use assessment to demonstrate school quality, but believed that assessment systems must measure highly valued outcomes such as deep learning. Thus, in the interim, they were cautious about the power of external, test-like assessments to evaluate schools fairly.

Consequently, there is both an empirical basis and a theoretical basis to support the proposition that differences in policy framework (associated with different jurisdictions and cultures) lead to differences in how assessment is conceived. Presumably, a low-stakes assessment regime would encourage the notion of assessment as *Improvement* rather than *Irrelevant*. Presumably also, a high-stakes assessment regime would lead to a weak

association between notions of *Improvement* and *School Accountability* and a greater focus on students as the people to be held accountable for their learning. One way to test these assumptions is to examine the impact on the structure of teachers' conceptions of assessment in other jurisdictions with a similar overall low-stakes policy framework for education. A second way would be to test the effect of very high-stakes policy priorities in a different jurisdiction (e.g., California).

The difference between primary and secondary education provides yet another possible explanation for the effect of context on teachers' conceptions of assessment. In many Western countries or states (e.g., England, Ireland, Scotland, Iceland, Greece, Portugal, New Zealand, and Queensland [Poulson, 2001; Vlaardingerbroek, & Taylor, 2003]) primary-school teachers are generalists, responsible for many subjects, and do not hold specialist or discipline-based degrees. In contrast, secondary school teachers tend to have specialist training (either a first degree in a discipline or special emphasis in their teaching qualification), which is weakly associated with better teaching practice (Floden & Meniketti, 2005). Furthermore, in contrast to secondary schooling, primary school is rarely treated as the terminal stage of education and so high-stakes assessment for certification or selection purposes do not frequently exist; with China as a notable exception (Gang, 1996). While testing for accountability or monitoring purposes may take place in primary schooling years (Levin, 2001), these assessments may not always hold consequences for individual children, being intended for evaluation of the school and not the student; with the notable exception of the English Key Stage testing system (Whetton, 2009) and various American school districts and states (Madaus, Russell, & Higgins, 2009). In contrast, high-stakes assessment (including external, public examinations) is deeply embedded in the secondary years of schooling. These assessments serve certification or qualifications requirements and may be administered continuously as part of school-based, internally assessed qualifications (as is the case in New Zealand and Queensland) or at the end of school-years in public, externally-administered examinations. Note that these structural differences are confounded by the increasing responsibility we expect students to take for their learning and lives as they develop through the teenage years in secondary school. Hence, we might expect, especially in jurisdictions where teachers act as in-school assessors or examiners for the qualifications system, that secondary teachers would place a greater emphasis on *Student Accountability* as the purpose of assessment.

1.3 Research context

The context for this research was the state of Queensland, Australia. At the time of this research, Queensland, like New Zealand, had (1) a low-stakes environment for educational assessment, (2) an outcomes-based curriculum framework, (3) restricted use of mandatory national testing, and (4) a highly-skilled teaching force.

Queensland has different assessment policies for primary (Years 1-7) and secondary (Years 8-12) levels of schooling. From Years 1 to 10 Queensland is an "assessment-free zone", a term used by the chair of the Queensland Education Department Assessment and Reporting Task to describe these years (C. Wyatt-Smith, personal communication, 21 February, 2002). Specifically, there were no common achievement standards or government-mandated common assessments in the primary and junior secondary years of schooling in Queensland (i.e., up to the 11th year of schooling). Queensland schools, at the time of this study, participated in federally funded tests, albeit not nationally standardised, for system-wide monitoring of literacy and numeracy at Years 3, 5, and 7. At the time of this study, these assessments were administered late in the school year with results being reported to schools at the start of the following school year, so their immediate impact as an accountability measure may have been minimal.

It is only in the senior secondary years (i.e., 11 and 12) that there is a rigorous system of externally moderated school-based assessment (which includes the application of state-wide standards). In Queensland, secondary teachers design and implement in-school assessment which relies heavily on socially-moderated teacher judgements as the basis for certification. Likewise, Queensland secondary teachers are largely subject specialists, while primary teachers are generalists. Hence, the difference of policy and qualifications between primary and secondary teachers in Queensland may well contribute to differences in how teachers conceive of assessment. Because many secondary-school teachers, being teacher-assessors for the qualifications systems, teach classes in both lower and upper secondary, it is highly likely that there would be a backwash effect from the qualifications system on teachers' conceptions of assessment, even for teachers with classes at Years 8 to 10. Assuming that the Year level taken by teachers affects their conceptions of assessment, we would expect to find *non*-invariance in the structure of teachers' conceptions of assessment and, more specifically, greater emphasis on student accountability among secondary school teachers.

1.3. Research purpose & questions

Within the context of evaluating potential assessment reforms in Queensland, two contextual factors that could affect conceptions of assessment presented themselves. One was the difference between Queensland and New Zealand. The other was the primary-secondary difference. Our purpose was to explore the effect of these two factors on conceptions of assessment in order to understand the relationship between societal context and teachers' conceptions.

This led to four specific research questions.

1. Could we model the four conceptions of assessment across Queensland primary and secondary teachers?
2. Is the model of assessment conceptions statistically equivalent between primary and secondary teachers?
3. Do the conceptions mean scores differ between primary and secondary teachers?
4. What are the conceptions of assessment among Queensland primary and secondary teachers?

2. Method

A questionnaire-based survey of teachers' attitudes, beliefs and practices in the areas of curriculum, pedagogy and assessment was conducted in 2003. The survey involved both primary and secondary government schools, providing a quasi-experimental design for examining the impact of jurisdiction and level. The questionnaire captured information about teachers' practices and conceptions of assessment (the subject of this paper), teaching (Author1, Author2, & Author3, 2009a), learning (Author1, Author2, & Author3, 2008), and curriculum (Author1, Author2, & Author3, 2009b).

2.1. Participants

The questionnaire was administered in November 2003 to all teachers in 92 state schools. Initial contact was made with schools to confirm their participation. Of the 3,223 questionnaires dispatched, 1,525 were returned, giving a gross teacher response rate of 47.3%. However, the true teacher response rate likely was higher since the sampling unit was the school and no responses at all were received from nine schools. We inferred that teachers in those schools did not have the opportunity to participate. The adjusted response rate based only on those 83 schools that participated (2,891 questionnaires dispatched) is 52.8%. For all

participants, the classes they taught were predominantly of students between Years 1 to 10, which was the intended population. Thus, it can be said that the data were provided by teachers not working in a high-stakes qualifications system. Participants who failed to supply at least 90% of answers to the survey were dropped from the analysis and any missing data from the remaining participants were imputed with the expectation maximization (EM) procedure (Dempster, Laird, & Rubin, 1977). The EM procedure calculates the observed item means, standard deviations, and covariance matrix from a set of variables and then, using an iterative maximum likelihood estimation technique, imputes missing values which produce new item means, standard deviations, and covariance matrix that are as close as possible to the original values. Statistically non-significant imputations typically result in item means and standard deviations that differ only at the third decimal place. As a consequence, valid data were obtained from 784 primary teachers and 614 secondary teachers.

The teachers were generally experienced (years of teaching M=13.63, SD=9.53) with only 23% having less than five years experience. Most had four to five years training (65%), with only 13% having less than four years training. Nearly three-quarters had either a four-year degree (52%) or a graduate diploma (21%), with majority of the balance having a three-year diploma (9%) or three-year degree (10%). Fully three-quarters held teacher positions, only 11% had leadership roles (i.e., principal, deputy principal, or head of department), and the remainder were specialist teachers (13%). On average the teachers had been in their current school at the time of the survey 5.51 years (SD=5.20), with as many as 37% having been in the school for only one or two years. Differences in the distribution of teachers by primary or secondary level were statistically not significant. This suggests that statistically significant differences in models or factor cannot be attributed to differences in teacher characteristics.

Teachers in Queensland state schools are employed and assigned centrally to schools and there is a high mobility rate across the large and diverse state. On the other hand, there is very little crossover of teachers between primary and secondary levels, even in schools that spanned both. Therefore, with this sample it is possible to detect whether there are differences based on the age of students (i.e., primary or secondary). Because of the high mobility, and large sample size, it is unlikely that there is any assortment at schools by sex, age, educational, or ethnicity, and, to the extent that we could test, the participants were representative of population of teachers employed in Queensland state schools.

2.2. Instrument

The instrument used was the abridged, 27-item *Conceptions of Assessment Inventory* (CoA-IIIA) (Author1, 2006), which was embedded in the larger questionnaire. The response scale for the items is a positively-packed agreement rating scale; that is, two negative options (i.e., mostly disagree and strongly disagree) and four positive options (i.e., slightly, moderately, mostly, and strongly agree) (Author1, 2004b). A skewed response scale is 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.

The CoA-IIIA elicits teacher self-ratings for the four inter-correlated conceptions of assessment discussed (i.e., *Improvement*, *Student Accountability, School Accountability* and *Irrelevant*). Note that two of these factors are hierarchical (i.e., *Improvement* contains four sub-factors each with three items; *Irrelevant* has three sub-factors each with three items). The overall fit of the 27-item nine-factor hierarchical model for the original 525 New Zealand primary school teacher sample was good (N=525, χ^2=841.02, df=311; χ^2/df=2.70, p=.10; gamma hat =.93; RMSEA =.057) (Author1, 2008). The standardized Cronbach alpha scale reliabilities were acceptable for the short scales and good for the longer scales (i.e.,

Improvement k=12, α=.85; *Student Accountability k*=3, α=.66; *School Accountability k*=3, α=.79; and *Irrelevant k*=9, α=.76). Among New Zealand primary teachers, the factor inter-correlations were weak to moderate among *Improvement*, *Student Accountability*, and *School Accountability* (range of *r*: .21 to .48), while the *Irrelevant* factor had quite variable inter-correlations (*Improvement r*=-.77; *Student Accountability r*=.35, *School Accountability r*=-.13). The average agreement score among New Zealand primary teachers for each factor ranged from below slightly agree to moderately agree (*Improvement*: *M*=4.08, *SD*=.70; *Student Accountability*: *M*=3.53, *SD*=.95; *School Accountability*: *M*=2.68, *SD*=1.01; *Irrelevant*: *M*=2.90, *SD*=.68); note all mean differences had large effect sizes (*d*>.60 [Hattie, 2009]), except between *School Accountability* and *Irrelevant* which was small (*d*=.26).

2.3. Analysis

Each research question has its own analytic characteristics.

RQ 1. Modeling Queensland teachers' conceptions of assessment

To answer our first research question we used confirmatory factor analysis, which is appropriate once a model has already been developed and tested with other data. In confirmatory factor analysis the model is specified, whereas in exploratory factor analysis it is developed from the data. The model was hierarchical; that is, there were first and second-order factors. The observed variables (the 27 items) are modelled as being caused by nine first-order factors, and seven of which are, in turn, caused by two second-order factors– together making the four conceptions of assessment. The four conceptions are correlated.

Confirmatory factor analysis tests the fit of this model to the data (Klem, 2000; Hoyle, 1995; Thompson, 2000). There are many measures to assess the fit of a model to the data. In line with current practice (Cheung & Rensvold, 2002; Fan & Sivo, 2007; Marsh, Hau, & Wen, 2004; Vandenberg & Lance, 2000) our criteria for fit were models with statistically nonsignificant χ^2 per *df*, gamma hat >.90, and root mean square errors of approximation (RMSEA) and standardized root mean residuals (SRMR) <.08. Models that met these criteria were not rejected.

When conducting confirmatory factor analysis we expect differences in fit and parameter values as a result of statistical variation and sometimes as a result of underlying differences in the two populations. But it may also be the case that the model is not admissible. Author1 (2006) reported that the model used to describe primary school teachers' conceptions of assessment in New Zealand fitted the responses of Queensland primary school teachers, but was inadmissible for Queensland secondary teachers. A model may be inadmissible for a number of reasons (Chen, Bollen, Paxton, Curran, & Kirby, 2001; Gerbing & Anderson, 1987), including small sample size which does not apply in this study.

A model may be inadmissible simply because, for a different population, the structure is mis-specified (i.e., has different paths). Modification indices identify paths, which, if added to the model, would improve fit and may resolve inadmissibility (Byrne, Shavelson, & Muthen, 1989). When using modification indices to respecify a model, the modifications must be theoretically defensible and, ideally, should be tested on an independent sample (MacCallum, 1995; Maruyama, 1998). In this study, since the original New Zealand model had already been shown to fit the Queensland primary school teachers' responses, modification indices were used to identify the least possible changes to the original model to make it admissible and well-fitting for both primary and secondary teachers.

RQ 2. Examining model differences between groups

Once we have developed a model that fits the data we can then look at the second research question about group differences. We do this by breaking the sample into two groups (i.e., primary and secondary) and then examining the parameters of the model to determine whether the values for the two groups differ by more than chance. Note, that in a multi-group analysis, the path parameter values are allowed to differ for each group, but there is only one set of fit indices for the total model. Statistical equivalence means that the differences in the path parameter values differ only at the chance level. Equivalence of five sets of parameters is normally needed to make comparisons between groups (Cheung & Rensvold, 2002; Vandenberg & Lance, 2000):

(1) all paths are identical (i.e., configural equivalence),
(2) all loadings from 1st-order factors to items are equivalent,
(3) all intercepts of item loadings on 1st-order factors are equivalent,
(4) all loadings from 2nd-order factors to 1st-order factors are equivalent, and
(5) all covariances between inter-correlated factors are equivalent.

Some analysts have argued that equivalent pathways and factor to item regressions (i.e., configural and metric invariance) are sufficient to compare factor scores (McArdle, 2007). Further, when invariance is demonstrated across all these parameters, we can conclude that the groups are members of the same population (Cheung & Rensvold, 2002; Wu, Li, & Zumbo, 2007) and that all path values are, within chance, equivalent across groups.

To establish the equivalence of the model between the level and pilot participation groups, nested multi-group invariance analysis was used (Byrne, Shavelson, & Muthen, 1989). This involves constraining the model to be equivalent for each of the five parameter sets listed above, examining the fit statistics, and moving to the next test only if the fit criteria indicated that the parameter values were equivalent. Testing stops when a parameter is shown not to be equivalent. The equivalence of the pathways is accepted if the RMSEA for a multigroup model is $\leq .05$. As the model is progressively constrained to be equivalent across groups, the difference in the comparative fit index is compared to the value for the model immediately preceding the constraint; the ΔCFI should be $\leq .01$ (Cheung & Rensvold, 2002; Wu, Li, & Zumbo, 2007).

RQ 3. Examining conceptions mean scores between groups

Given that the same underlying conceptions exist in different groups, we would expect that any effect of context would be manifest in mean score differences for the four conceptions. Conception scores were the average of all items contributing to a conception; the items were scored 1 strongly disagree, 2 mostly disagree, 3 slightly agree, 4 moderately agree, 5 mostly agree, and 6 strongly agree. A MANOVA *F*-test was used to establish statistically significant differences between groups; though it should be noted with the large sample sizes, statistical significance will be easy to achieve. To establish the practical significance of differences, the difference in mean scores was calculated as Cohen's (1977) effect size (*d*). Hattie (2009) has shown that in education research values of *d* up to .20 are trivial, between .21 and .39 are small, between .40 and .59 are moderate, and >.60 are large.

RQ 4: Understanding differences in Queensland teachers' conceptions of assessment

The overall meaning of the model was determined by examining the factor inter-correlations, factor mean scores, and the strength and direction of regression weights between factors. These values indicate how strongly groups endorse each construct and how each construct is influenced by the other constructs. They also indicate how strongly and in which direction the various sub-beliefs contribute to the four main latent conceptions of assessment.

The overall pattern of these parameters is used to create a picture of how the various groups understand and respond to assessment.

3. Findings

3.1. RQ 1. Modeling Queensland teachers' conceptions of assessment

The problem of inadmissibility of the secondary teacher group was solved by introducing two new paths between first-order factors. Two paths – from the 'assessment improves student learning' factor to the 'assessment is inaccurate' factor and from the *Student Accountability* factor to 'the assessment describes learning' factor – were added. These were conceptually plausible. The capacity of assessment to improve student learning depends, in part, on its accuracy, and the capacity of assessment to make students accountable depends, in part, on its ability to meaningfully describe what students learn. This respecified model, having the same nine factors as the original New Zealand model and two additional paths, was admissible and had marginal fit (df=309; χ^2=2442.60; χ^2/df=7.91, p<.01; CFI=.844; gamma hat=.90; RMSEA=.070; SRMR=.087). By breaking the data set into two groups (i.e., primary and secondary—Figures 1 and 2) and allowing the one model to have different parameter values for each group, the fit of the model improved considerably (df=618; χ^2=2741.56; χ^2/df=4.44, p=.04; CFI=.846; gamma hat=.95; RMSEA=.050; SRMR=.089), indicating that the two groups had independent responses to the inventory.

INSERT Figures 1 and 2 about here

3.2. RQ 2. Examining model differences between groups

The respecified model was configurally invariant (RMSEA=.050). The regression weights from the first-order factors to items were equivalent (ΔCFI=.002), but the intercepts at the first-order factors (i.e., scalar invariance) were not (ΔCFI=.041). The differences in the inter-factor correlations were small (p>.05), except for the correlation between *Irrelevance* and *Student Accountability*, which was much stronger (p<.001) for the primary teacher group. We conclude that the two groups of teachers were drawn from different populations and that the model differences represent real differences in primary and secondary teachers' conceptions rather than vagaries of the model or the instrument. There was sufficient similarity in the model for the factor scores to be compared, but because there were statistically significant differences in the inter-factor correlations and pathway regression weights, the model has to be interpreted separately for each group.

3.3. RQ 3. Examining conceptions mean scores between groups

Table 1 shows the conceptions mean scores by level of schooling and the statistical and practical significance of the difference in mean scores. Two conceptions – *Improvement* and *Student Accountability* – had statistically significant different means across level, although the scale of difference was small to medium. Primary teachers were higher on *Improvement* , and secondary teachers were higher on *Student Accountability*. The differences for *Irrelevant* and *School Accountability* were trivial and within chance.

Insert Table 1 about here

3.4. RQ 4: Understanding Queensland teachers' conceptions of assessment

Each conception is interpreted separately and differences/similarities between primary and secondary teachers and to studies with New Zealand (Author1, 2008) and Hong Kong (Author1, et al., 2009) teachers are discussed.

3.4.1. *Improvement*

As in the New Zealand and Hong Kong models, the *Improvement* conception consisted of four first-order factors. In order for the model to fit, we needed a new path ($\beta = .29$ for both groups) from the 'students' learning' first-order factor to 'assessment is inaccurate' (one of the first-order factors defining *Irrelevant*). This suggests an inter-relationship between belief in the accuracy of assessment and its utility in providing information to students that will assist their learning, which is consistent with the literature. Teachers tended to associate improvement with keeping in mind, as advised by psychometric experts (e.g., Linn & Gronlund, 2000; McMillan, 2001b), that assessment judgments are partially contingent on the degree of error in an assessment. The mean score for *Improvement* was highest (albeit only moderately agree) for primary teachers, but not the highest for secondary teachers.

As in New Zealand, the *Improvement* was positively correlated with *School Accountability*, inversely correlated with *Irrelevant*, and the correlation with *Student Accountability* was not statistically different from zero. Note that the differences in correlations between the two groups were statistically not significant. In contrast, there was one significantly different correlation in the Hong Kong study (Author1, et al., 2009) between *Improvement* and *Student Accountability*, which was strongly positive, a difference attributed to the level of consequences given to assessment practices in Chinese societies. For Queensland teachers a belief in *Improvement* was the opposite of a belief in *Irrelevant*, which is a consistent and coherent belief system. Assessment for learning (*Improvement*) was positively associated with demonstrating the quality of schools (*School Accountability*), while it had no systematic relationship with *Student Accountability*.

3.4.2. *Irrelevant*

As in New Zealand and Hong Kong, the *Irrelevant* conception consisted of three first-order factors. Assessment was irrelevant when it was 'bad for teachers and students', when it was 'used but ignored', and when it 'contains error'. The mean score for both primary and secondary teachers were low (below slightly agree). Unlike New Zealand, the loadings onto 'assessment contains error' were much stronger for both groups of Queensland teachers, suggesting that these teachers were much more inclined to credit inaccuracy as the grounds for treating assessment as irrelevant. Note this result is quite different to the Hong Kong study in which the same items had weak negative loadings on irrelevance, suggesting Chinese teachers did not have to take inaccuracy into account. This finding should concern policy makers and administrators – the lower the quality of the assessment methods used, the more irrelevant teachers will find them, and rightly so in our view. Furthermore, of just as much concern are the 5-7% of Queensland teachers (>1.5 *SD* above *M*) who gave this conception a mean score ≥4.0, indicating that they believed assessment had little legitimate role in education. The provision of assessment instruments of high quality (i.e., sensitive, without bias, not highly inaccurate, and useful in a classroom context) is vital in persuading teachers that assessment is **not** irrelevant.

Like New Zealand and unlike Hong Kong, *Irrelevant* was correlated with *Student Accountability*, showing that the more teachers believed in the irrelevance of assessment, the more they rejected its use in making students accountable. Teachers in both groups were more likely to see assessment as *Irrelevant* when they considered individual students, the characterization being much more substantial for primary teachers. The strength of the

correlation was significantly higher (z=7.69) for primary teachers than secondary teachers. It would appear that Queensland teachers, and especially primary teachers, are as child-centred as New Zealand teachers, considering assessment to be bad for children. The paths here indicate part of the reason for this thinking; assessment is bad if it has a student accountability focus and is inaccurate rather than aiming for improved learning, or even having a school accountability focus.

3.4.3. *School Accountability*

As in the New Zealand and Hong Kong models, this conception consisted of three strongly loading items that focused on the use of assessment to determine the quality of a school. The mean scores for both primary and secondary teachers were the lowest (≈2.7 on a 1 to 6 scale). *School Accountability* was moderately correlated with *Student Accountability* and *Improvement*, but not with *Irrelevant*; a pattern consistent with both Hong Kong and New Zealand studies. It should be noted that the difference between the two groups in this last correlation was statistically significant (z=2.04) but, since the absolute value was statistically non-significant in both groups, this difference is not discussed further.

In other words, teachers who viewed the use of assessment as a school accountability mechanism tended also to view the use of assessment as a student accountability mechanism. At the same time they conceived of assessment as improving the quality of teaching and learning. Accountability at the school level, assessing students, and improvement were intertwined rather than juxtaposed. Thus, teachers did not exhibit the simplistic notion of formative assessment good, summative assessment bad. There was a complex interrelationship between these potentially contrary purposes, suggesting that allowing these tensions to exist simultaneously is a necessary precondition for successful professional development in assessment. Nonetheless, the teachers were opposed to using assessment for evaluating schools—a matter of some concern to administrators, though the positive correlation with *Improvement* suggests that teachers may believe good schools improve learning.

3.4.4. *Student Accountability*

Again like New Zealand and Hong Kong, the student accountability conception consisted of the three items focused on grading, categorising, and evaluating students. The mean scores for this conception were relatively strong (i.e., approaching 4.00). In order to achieve model fit we added a path to the ‘assessment describes what students know’ first-order factor. Teachers, most especially those in secondary schools, accepted that assessment makes a student accountable through grades, categories, or certificates because it provides information about what a student has learnt, including higher-order thinking skills. It seems to us that acceptance of the claim that assessment provides this type of information is a necessary precursor for making students accountable.

Student Accountability was moderately correlated with *School Accountability* – if schools are held accountable through assessment, those results are generated by individual students who were assessed and, thus, the conceptions are logically connected. However, making students accountable was not related to *Improvement* (i.e., correlations were not statistically significant), suggesting that assessments of student learning were not systematically associated with improvement — perhaps the teachers had experience of assessments that were divorced from real learning? Further, there was a positive correlation with *Irrelevant*, suggesting that teachers associated grading of students as irrelevant, rather than as a means of improvement. The pattern of results here are highly similar to the previously reported New Zealand study and quite different to the Hong Kong study.

4. Discussion

This study has answered four research questions about modelling and understanding Queensland teachers' conceptions of assessment and their relationship to their level of teaching and compared the results to teachers from New Zealand. Evidence was found that four conceptions of assessment (i.e., *Improvement*, *Irrelevant*, *School Accountability*, and *Student Accountability*) could be used to describe how practicing teachers understood assessment, specifically by examining the purpose inter-correlations, their contributing factors, and their mean scores. This study provided support for using the *Teachers' Conceptions of Assessment Inventory* to research practicing teachers' conceptions of assessment.

4.1 Policy impact on conceptions

The quality of measurement was such that differences between groups were most likely due to real population differences rather than chance artefacts in responding to the questionnaire. Fundamentally, Queensland primary and secondary teachers' conceptions of assessment were more alike than different. However, consistent with the notion that policy shapes conceptions, small, statistically significant, and important differences were noted around the use of assessment to make students accountable and the use of assessment to improve teaching. The general culture of Queensland expresses itself in two relevant policies: (1) high-stakes examinations only take place in senior secondary school, and (2) Queensland education should prepare its citizens for the future (Queensland Government, 2001). The finding about school level is confounded with normal expectations that secondary students take more responsibility and can be held accountable for their performance. Nonetheless, the views expressed by primary and secondary teachers respectively appear to align with these cultural priorities, whether they are developmental or policy-led. Furthermore, the results were very similar to those of New Zealand teachers, a jurisdiction with very similar assessment policies, and very different to those of Hong Kong teachers, a jurisdiction with very different policies and priorities.

Notwithstanding higher mean scores for *Student Accountability* and low mean scores for *School Accountability*, there appears to be a potentially positive latent relationship in the inter-correlations among these factors. The notion of assessment for *Improvement* was associated with *School Accountability* but not systematically related to *Student Accountability*. Teachers appeared to be willing to take responsibility for improving school outcomes and quality and fulfilling their professional accountabilities, while rejecting the notion that assessment should focus on students. These data revealed that teachers did not have an anti-assessment mentality—rather teachers showed a willingness to integrate assessment into their professional duties of improved teaching and learning, tempered with caution about the quality and usefulness of the assessment resources being used to make students and schools accountable. If the assessments were valid, informative, and connected to classroom improvement, then the current conceptions of teachers suggested that such assessments would be welcomed. However, if the assessments were seen as unfairly punitive of children, of dubious quality, or lacking in power to improve classroom learning, then they would be considered irrelevant. The message is not so much that the teachers were assessment illiterate, but rather that their conceptions were consistent with the requirements of validity.

4.2 Implications for policy development and teacher education

It is interesting for teacher education to examine, not just the degree of endorsement teachers have for each of the four purposes, but rather how teachers relate those purposes to

each other. The TCoA inventory and this factor analytic model permits identification of strength of opinion, its constituent parts, and its relationship with other beliefs. In this way, a sophisticated analysis of how various policy conditions impact on teacher understanding could be undertaken. If, as Pajares (1992) has argued, teacher beliefs arise out of schooling experiences as students, then the use of the TCoA inventory with prospective teachers may be a useful adjunct to teacher education. The inventory appears to be both an efficient and valid means of establishing a baseline early in student preparation and for monitoring changes in conceptions of assessment in response to explicit teaching about assessment and teaching practicum experiences.

Inasmuch as governments rely on teachers to carry out assessment innovations and policies, the beliefs, values, and attitudes teachers have about assessment will influence and shape their implementation of such initiatives. We are told that teachers need more assessment literacy, but these results instead suggest, in our view, that policy makers, professional developers, teacher educators, and administrators may have failed to persuade teachers that the currently available assessment systems provide informative, valid, and improving effects. This failure may lie in the inadequacies of the assessments or in the failure of assessment systems to generate usable information in a timely manner. Likewise, the failure may lie in the act of being externally reviewed; external surveillance itself (with or without consequences) may raise concerns amongst teachers that invalidate the assessment system in their minds.

Professional development and teacher preparation for participants holding the mix of views about assessment found in this sample of Queensland teachers should seek to take advantage of teachers' commitment to assessment as a relevant means of improved teaching and learning. The antipathy for holding both schools and individuals accountable through assessment may not be so much a desire to escape responsibility, but rather a rational rejection of poor-quality assessment systems that have unjust stakes or consequences for schools and/or learners. There is not much point making teachers or students accountable if the assessments being used are not defensibly aligned with teaching, learning, and curriculum, if they are not timely and rich in their feedback to the teacher, or if they are patently inaccurate or unfair. Perhaps the point of teachers' assessment literacy is not so much a problem of teacher knowledge or teacher thinking, but rather one of limited access to good assessment design and use.

Like New Zealand, Queensland teachers in this study work in a context of low-stakes assessments designed to improve classroom practices or inform central agencies about the quality of the system. In Queensland, central agencies administer annual tests which cannot effectively hold students or teachers accountable, since the national tests are not aligned to the classroom curriculum, nor are they timely in their reporting to schools. However, New Zealand teachers since 2003 have been given a wide array of voluntary-use, non-centrally controlled, diagnostic, quick feedback, teacher-controlled assessments that have allowed teachers to improve learning in a self-managed manner (Hattie & Author1, 2008; Hattie, Author1, & Keegan, 2003; Hattie, Author1, Ward, Irving, & Keegan, 2006). In that context, Author1 (2008) reported that New Zealand primary teachers actually associated the use of assessment for improvement with school accountability. Our message to policy makers is that radically different, low-stakes, richly informative, highly aligned assessments may be needed to engender a robust conception among teachers that assessment improves teaching and learning and that it can be used to demonstrate accountability.

4.2 Implications for future research

The results reported here flow from interpreting survey responses. Corroborating evidence for these interpretations through other means (e.g., interviews, observations of

practice, and assessment samples) is needed to support the interpretations offered here. An interview study in New Zealand used the four purpose framework to capture the complexity of teacher assessment beliefs (Harris & Author1, 2009a), although the interview and questionnaire responses were at best complementary (Harris & Author1, 2009b). Whereas Author1 (2002) reported that teacher conception mean scores were statistically invariant by the type of practices teachers associated with the term assessment, it seems plausible that specifying the type of assessment to be evaluated (e.g., when answering this questionnaire think of high-stakes national testing vs. think of low-stakes classroom assessment) would be a good approach to experimentally validating interpretations.

The model reported here is conceptually similar to that used to analyse New Zealand and Hong Kong primary and secondary teachers' responses (Author1, 2008; Author1, et al., 2009). The differences in the Queensland model suggest there may be cultural or societal effects on teacher responses to the inventory. Nonetheless, given the fit of the Queensland model based on responses to the same inventory, its use as a research tool in other jurisdictions appears warranted. We would expect that teachers in high-stakes assessment jurisdictions (e.g., the United States or England) where the assessment systems are less aligned to classroom learning, where the assessments do not provide rich and timely feedback, where the assessment system has greater stakes or consequences, and where the assessments may not accurately or validly describe student learning, would have a different pattern of agreement with these conceptions. In such contexts, teachers might be expected to disagree that assessment improves teaching and learning; they may still see it as relevant, but it probably would be rejected as a means of school accountability. We look forward to such studies and suggest that the current instrument would be of great value in such research.

Acknowledgements

Funding for this research came from the Queensland Department of Education and the Arts, Assessment & New Basics Branch and from The University of Auckland Research Office. The opinions expressed here are those of the authors only.

References

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

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

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher, 36*(5), 258-267.

Author1 (2002). [details removed for peer review]

Author1 (2004a). [details removed for peer review]

Author1 (2004b). [details removed for peer review]

Author1 (2006). [details removed for peer review]

Author1 (2008). [details removed for peer review]

Author1 (2009). [details removed for peer review]

Author1, Author2, & Author3 (2008). [details removed for peer review]

Author1, Author2, & Author3 (2009a). [details removed for peer review]

Author1, Author2, & Author3 (2009b). [details removed for peer review]

Author1, Kennedy, K. J., Fok, P. K., Chan, J. K. S., & Yu, W. M. (2009). [details removed for peer review]

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, IL: NSSE & University of Chicago Press.

Butterfield, S., Williams, A., & Marr, A. (1999). Talking about assessment: mentor-student dialogues about pupil assessment in initial teacher training. *Assessment in Education: Principles, Policy & Practice, 6*(2), 225-246.

Byrne, B. M., Shavelson, R. J., & Muthen, B. (1989). Testing for the equivalence of factor covariance and mean structures –- the issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456-466.

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.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233–255.

Choi, C. C. (1999). Public examinations in Hong Kong. *Assessment in Education: Policy, Principles and Practice, 6*(3), 405-417.

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.

Cohen, J. (1977). *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press

Cooper, P., & Davies, C. (1993). The impact of national curriculum assessment arrangements on English teachers' thinking and classroom practice in English secondary schools. *Teaching & Teacher Education, 9*, 559-570.

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

Crooks, T. J. (1990). Overview of national monitoring in education. In A. C. Croft (Ed.), *Proceedings of the Conference on National Monitoring in Education* (pp. 5-21). Wellington, NZ: NZCER.

Darling-Hammond, L. (2003, February). Standards and Assessments: Where We Are and What We Need *Teachers College Record* Retrieved 2 August, 2005, from http://www.tcrecord.org

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B, 39*(1), 1-38.

English, F. W. (1988). The utility of the camera in qualitative inquiry. *Educational Researcher, 17*(4), 8-15.

Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research, 42*(3), 509–529.

Firestone, W. A., Mayrowetz, D., & Fairman, J. (1998). Performance-based assessment and instructional change: The effects of testing in Maine and Maryland. *Educational Evaluation and Policy Analysis, 20*(2), 95-113.

Floden, R. E., & Meniketti, M. (2005). Research on the effects of coursework in the arts and sciences and in the foundations of education. In M. Cochran-Smith & K. M. Zeichner (Eds.), *Studying teacher education: The report of the AERA Panel on Research and Teacher Education* (pp. 261-308). Mahwah, NJ: LEA.

Gang, W. (1996). Educational assessment in China. *Assessment in Education: Principles, Policy & Practice, 3*(1), 75-88.

Gerbing, D. W., & Anderson, J. C. (1987). Improper solutions in the analysis of covariance structures: Their interpretability and a comparison of alternate respecifications. *Psychometrika, 52*(1), 99–111.

Gipps, C., Brown, M., McCallum, B., & McAlister, S. (1995). *Intuition or evidence? Teachers and national assessment of seven-year-olds.* Buckingham, UK: Open University Press.

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*(1), 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., & Author1 (2009a). [details removed for peer review]

Harris, L. R., & Author1 (2009b). [details removed for peer review]

Hattie, J. (2009). *Visible learning: A synthesis of meta-analyses in education*. London: Routledge.

Hattie, J. A. C., & Author1 (2008). [details removed for peer review]

Hattie, J. A. C., Author1, & Keegan, P. J. (2003). [details removed for peer review]

Hattie, J. A., Author1, Ward, L., Irving, S. E., & Keegan, P. J. (2006). [details removed for peer review]

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. *The Journal of Educational Research, 93*, 276-286.

Klem, L. (2000). Structural equation modeling. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and Understanding More Multivariate Statistics* (pp. 227–260). Washington, DC: APA.

Koretz, D. (2002). Comment. In D. Ravitch (Ed.), *Brookings Papers on Education Policy: 2002.* (pp. 315-323).

Levin, B. (2001). *Reforming Education: From Origins to Outcomes*. London: RoutledgeFalmer.

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

Linn, R. L., & Gronlund, N. E. (2000). *Measurement and assessment in teaching.* (8th ed.). Upper Saddle River, NJ: Merrill/Prentice Hall.

MacCallum, R. C. (1995). Model specification: Procedures, strategies, and related issues. In R. H. Hoyle (Ed.), *Structural Equation Modeling: Concepts, Issues, and Applications* (pp. 16–36). Thousand Oaks, CA: Sage.

Madaus, G., Russell, M., & Higgins, J. (2009). *The paradoxes of high stakes testing: How they affect students, their parents, teachers, prinicpals, schools, and society*. Charlotte, NC: Information Age Publishing.

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

Maruyama, G. M. (1998). *Basics of structural equation modeling*. Thousand Oaks, CA: Sage.

McArdle, J. J. (2007). Five steps in the structural factor analysis of longitudinal data. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (pp. 99–130). Mahwah, NJ: LEA.

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

McMillan, J. H. (2001b). Classroom assessment: Principles and practice for effective instruction (2nd ed.). Boston, MA,: Allyn & Bacon.

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

Mehrens, W. A., & Lehmann, I. J. (1984). *Measurement and evaluation in education and psychology* (3rd ed.). New York, NY: Holt, Rinehart and Winston.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). Old Tappan, NJ: MacMillan.

Ministry of Education. (1994). *Assessment: Policy to Practice*. 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.

Neill, M. (1997). *Testing our children: A report card on state assessment systems*. Cambridge, MA: FairTest.

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, CA: 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.

Philipp, R. A., Flores, A., Sowder, J. T., & Schappelle, B. P. (1994). Conceptions and practices of extraordinary mathematics teachers. *Journal of Mathematical Behavior, 13*, 155-180.

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

Porter, A., & Chester, M. (2002). Building a high-quality assessment and accountability program: The Philadelphia example. In D. Ravitch (Ed.), *Brookings Papers on Education Policy: 2002* (pp. 285-337.).

Poulson, L. (2001). Paradigm lost? Subject knowledge, primary teachers and education policy. *British Journal of Educational Studies, 49*(1), 40-55.

Queensland Government. (2001). *Years 1-10 Curriculum Framework for Education Queensland Schools: Policy and Guidelines*. Brisbane, Qld.: The State of Queensland.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement, and instruction* (pp. 37-75). Boston, MA: Kluwer Academic Publishers.

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

Scriven, M. (1991). Beyond formative and summative evaluation. In M. W. McLaughlin & D. C. Phillips (Eds.), *Evaluation & education: At quarter century* (Vol. 90, Part II, pp. 19-64). Chicago, Il.: NSSE.

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.

Smith, M. L., Heinecke, W., & Noble, A. J. (1999). Assessment policy and political spectacle. *Teachers College Record, 101*(2), 157-191.

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.

Thompson, B. (2000). Ten commandments of structural equation modeling. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and Understanding More Multivariate Statistics* (pp. 261–283). Washington, DC: APA.

Thorndike, R. M. (1997). *Measurement and evaluation in psychology and education.* (6th ed.). Upper Saddle River, NJ: Prentice-Hall.

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.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(4), 4–70.

Vlaardingerbroek, B., & Taylor, T. G. N. (2003). Teacher education variables as correlates of primary science ratings in thirteen TIMSS systems. *International Journal of Educational Development, 23*, 429-438.

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.

Whetton, C. (2009). A brief history of a testing time: National curriculum assessment in England 1989-2008. *Educational Research, 51*(2), 137-159.

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research & Evaluation, 12*(3), Available online: http://pareonline.net/getvn.asp?v=12&n=13.

Table 1

Conceptions of Assessment Mean Scores and Differences by Level of Teaching

Purposes of Assessment	*M*	*SD*	*M*	*SD*	*F*	*p*	*d*
Level of Teaching	Primary		Secondary		Difference		
Improvement	4.00	.67	3.84	.66	19.67	.00	.24
Irrelevant	2.89	.76	2.87	.69	.19	.66	.02
School Accountability	2.74	1.12	2.69	1.10	.81	.37	.05
Student Accountability	3.64	.76	3.93	.75	51.93	.00	-.39

Note. Negative values for *d* indicate Primary is lower, positive values indicate Primary is higher. *N* = 784 Primary; 614 Secondary; 850.

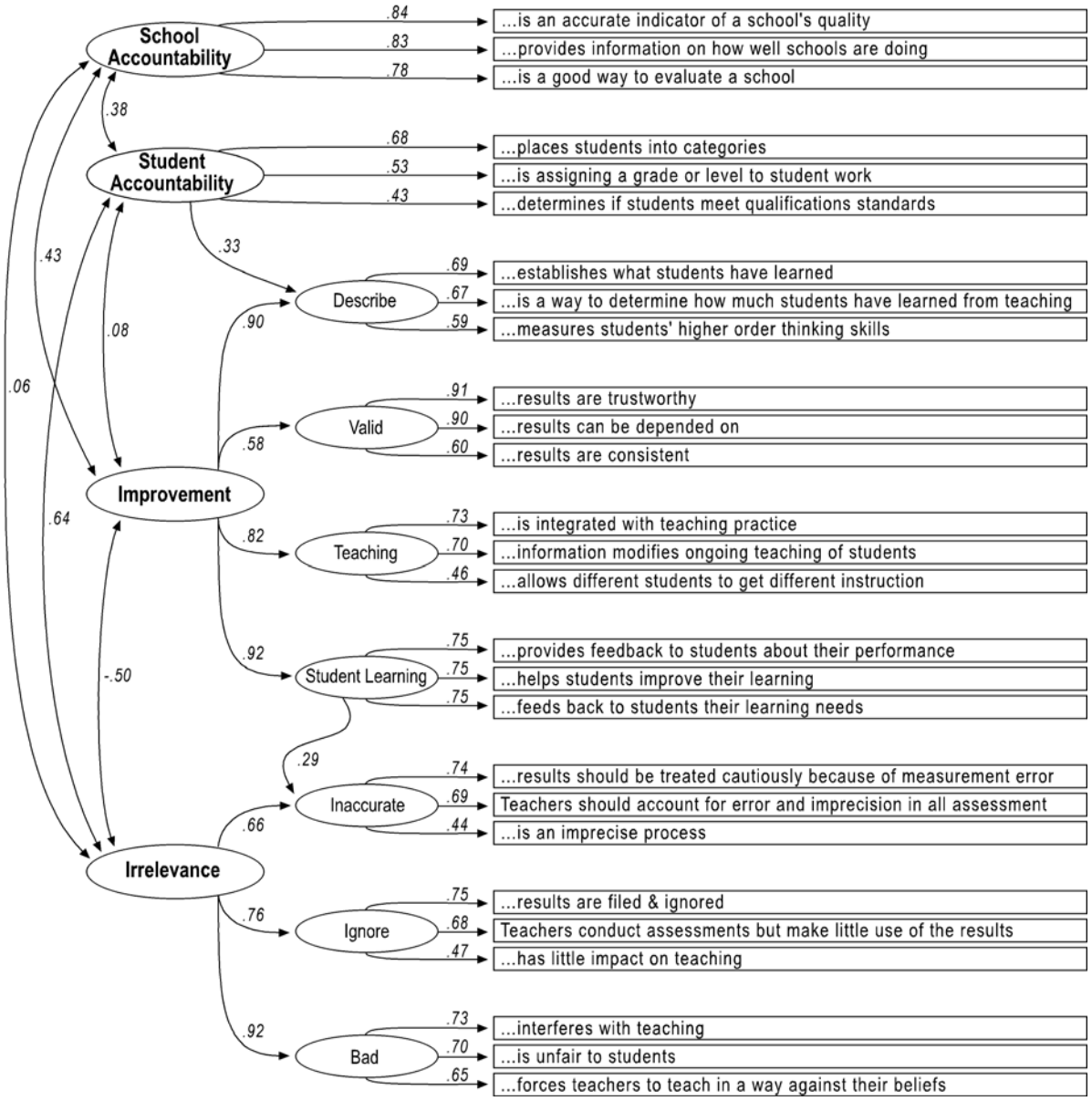

Figure 1. Queensland primary teachers' conceptions of assessment

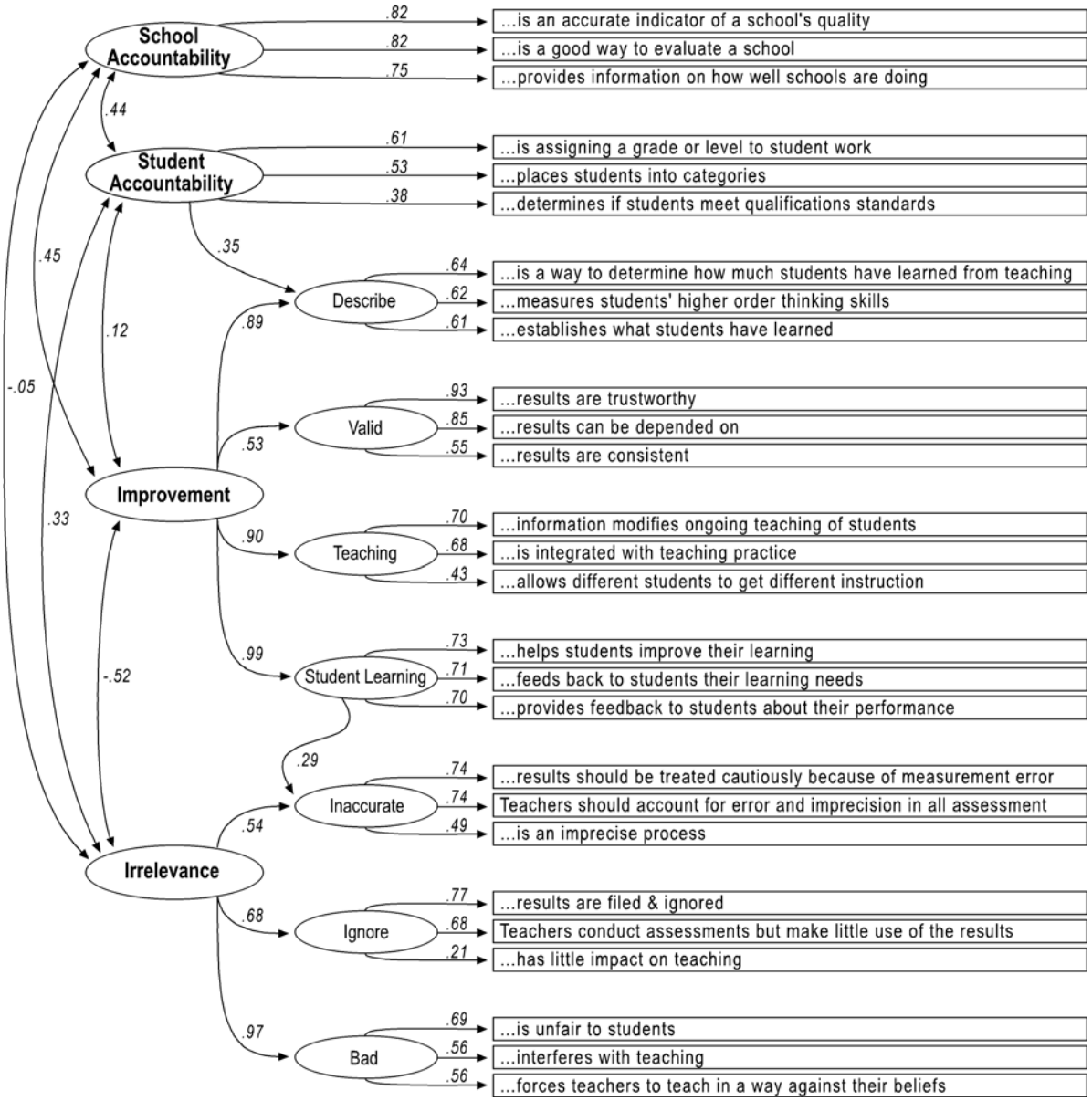

Figure 2. Queensland secondary teachers' conceptions of assessment