Teachers’ conceptions of assessment in Chinese contexts: A tripartite model of accountability, improvement, and irrelevance

Gavin T.L. Brown a,*, Sammy K.F. Hui b, Flora W.M. Yu b, Kerry J. Kennedy b

a The University of Auckland, New Zealand
b The Hong Kong Institute of Education, China

A R T I C L E   I N F O

Article history:
Received 9 September 2010
Received in revised form 20 September 2011
Accepted 10 October 2011
Available online 8 November 2011

Keywords:
Beliefs
Values
Attitudes
Teachers
Educational assessment
Confirmatory factor analysis
Multi-group invariance testing
Survey research

A B S T R A C T

The beliefs teachers have about assessment influence classroom practices and reflect cultural and societal differences. This paper reports the development of a new self-report inventory to examine beliefs teachers in Hong Kong and southern China contexts have about the nature and purpose of assessment. A statistically equivalent model for Hong Kong and southern China teachers had three factors (i.e., improvement, accountability, and irrelevance). The Chinese teachers very strongly associated accountability with improvement (r = .80). This is consistent with the Chinese tradition and policy of using examinations to drive teaching quality and student learning and as a force for merit based decisions. Small differences between the two groups of teachers are consistent with assessment policy differences in the two jurisdictions.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

Current educational policy reform movements attempt to reduce the negative consequences of highly selective educational assessments by placing a greater emphasis on using assessment to inform teaching and learning improvements through actively involving learners in guiding learning development (i.e., assessment for learning—Assessment Reform Group, 1999). Assessment for learning policy initiatives, while arising especially in England as a response to high-stakes national testing (Stobart, 2006), are being exported throughout the world, including Hong Kong (Berry & Adamson, 2011; Kennedy & Lee, 2008). The policy reform places much greater responsibility on the teacher to effectively and professionally use assessment information to modify teaching practice and provide corrective feedback to students who are expected to be actively engaged learners (Leahy, Lyon, Thompson, & Willam, 2005). This pressure requires that teachers understand the rationale for improvement-oriented assessment and have the knowledge, skills, and resources to put this policy reform into action. In societies with high-stakes, public examination systems that are used to select students for progressively reducing educational opportunities (e.g., China and Hong Kong—Kennedy, 2007), this policy may fail through lack of teacher cooperation, knowledge, or belief in the proposed new usage of assessment. Hence, it becomes critical, not just for the success of policy initiatives, but also for the development of pre-service and in-service teacher education, to understand how assessment is understood by teachers.

* Corresponding author. Tel.: +64 9 623 8890x48602.
E-mail address: gt.brown@auckland.ac.nz (Gavin T.L. Brown).
1.1. Teacher conceptions of assessment

Teachers’ opinions, attitudes, and beliefs (i.e., conceptions—Thompson, 1992) play an important part in mediating how educational reforms are implemented in schools and classrooms (Richardson & Placier, 2001). Explicit attention to teachers’ conceptions of the purposes of assessment and their practices of assessment is important since much educational policy related to assessment is implemented by and through school teachers. While there are many specific uses to which assessments can be put (e.g., selection, certification, diagnosis, transfer, licencing, monitoring, etc. described in Newton, 2007), we suggest that these can be reduced to four major conceptions of assessment, three of which may loosely be categorised as ‘purposes’ and one as an ‘anti-purpose’ (Brown, 2008). Three major purposes for assessment thread the scholarly literature (e.g., Black, 1998; Heaton, 1975; National Research Council, 2001; Torrance & Pryor, 1998; 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). A self-administered inventory has been developed, the Teacher Conceptions of Assessment (TCoA) inventory (Brown, 2001–2003), which generates inter-correlated agreement scores for these four constructs. Development was carried out in New Zealand with primary school teachers (Brown, 2004) and replicated with Queensland primary teachers (Brown, 2006).

There is a strong tension between accountability/evaluation uses and improvement-oriented purposes specifically related to educational assessment, testing, and examination. On the one hand, assessments are used to evaluate the quality of schools and teachers, as well as certify the learning of students, and, on the other hand, assessments are used to inform teachers, administrators, governments, parents, and students as to what aspects of learning have been mastered and what aspects need to be taught and learned next. The problem with this dual evaluative and improvement function is that it has been clearly shown that subordinates’ beliefs (e.g., teachers) become conformist when the views of an authority (e.g., principal or education authority) to whom individuals are accountable are known (Lerner & Tetlock, 1999). Simplistically, Hong Kong school authorities and parents subscribe to the notion that good schools generate high scores and grades on official examinations and qualifications. Teachers are aware that the quality and worth of their work arises from such outcomes. While not necessarily incompatible or mutually exclusive, whenever high-stakes socially or politically mandated consequences are attached to assessment results, it seems rational to expect the accountability purpose to dominate in the thinking of teachers.

1.2. Context effects on teacher conceptions of assessment

It is expected that differences in culture or society lead not only to differing policies, but also distinctive conceptions of practices or processes. We assume that conceptions are ecologically rational representations of the thought and practice traditions an individual experiences within a culture. While the United States and other English-speaking countries share common socio-cultural roots, there are considerable differences as to which educational policies are implemented and how they are carried out. 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. Based on survey studies using the TCoA inventory, it was found that teachers in New Zealand and Queensland, which have similar low-stakes assessment systems, had very similar conceptions of assessment, except for the notion of assessment being used to make schools accountable which was endorsed more by New Zealand primary teachers, a difference attributed to a greater emphasis on school self-management in New Zealand (Brown, Lake, & Matters, 2011). Hence, it would appear that how teachers understand and value the competing purposes of assessment is sensitive to policy priorities. However, in comparison to these reasonably similar societies, East Asian societies (e.g., Hong Kong, China, Japan, and Korea), while having similar cultural and policy practices to each other, are quite different to Western societies on many historical, linguistic, ethnic, and cultural levels. The policies adopted in those jurisdictions may reflect cultural differences as much as the normal variation of policy choices seen in the west.

This means that in societies with different assessment policies and practices, teachers’ conceptions of assessment should differ and their responses to the TCoA inventory, for example, should not fit well to a statistical model developed in a different culture. Lack of fit to a pre-existing model may suggest that the factors of the inventory may not be sufficient for use in other societies, even with high-quality translation/adaptation. Further, even if the same factors are present, it may be legitimate to expect the pattern and strength of the paths among the factors, let alone the strength of endorsement for each factor, would be statistically different for samples taken from quite different populations. In other words, while some factors may be stable across populations, we can legitimately expect the relationships among and within those factors to differ across societies.

1.3. Assessment in Chinese contexts

Both Hong Kong and China have long traditions of high-stakes examinations to select students for limited spots in higher levels of education or in higher-rated educational institutions. Indeed, there is at least 1000 years of history and social
support behind the use of public examinations as a selection tool in Chinese contexts (Paine, 1990). Cheung (2008) makes it clear that public examinations are necessary even in contemporary Hong Kong to prevent corruption and collusion in the selection of meritorious candidates for limited resources. Indeed, teachers within Confucian heritage societies appear to see frequent summative assessment and practice for formal examinations as a means of motivating effort and as a means of guiding instruction (Kennedy, Chan, Fok, & Yu, 2008). An additional important distinction in the function of assessment in Chinese contexts is that a good person is one who scores well because examination results reflect the quality and worth of the individual (China Civilisation Centre, 2007; Li, 2009; Pong & Chow, 2002). Thus, helping students score more on assessments increases not only their knowledge and performance, but also helps them become better people. This notion is operationalized in a cultivating (Gao & Watkins, 2001) response in which teachers help students to become better academically, attitudinally, and behaviourally. Hence, in the Chinese society of Hong Kong, high academic performance demonstrates perfection and superiority of character.

To exemplify these societally defined practices, consider the use of examinations in Hong Kong. The Hong Kong Attainment Test run in Pre-S1 helps classify feeder primary schools to the three attainment bands of secondary schooling. High-stakes examinations in Years 11 and 13 (i.e., HKCEE and HKALE) select a diminishing number of candidates for opportunities in the next level of schooling (e.g., 60% of HKCEE graduates win places in government funded 6th Form colleges and only 18% of the cohort obtain funded places in Hong Kong universities). In response to this examination culture, the Hong Kong Education Bureau (EDB) has promulgated a curriculum framework (Kennedy, 2005) and assessment tools and systems (details in Brown & Ngan, 2010) to assist schooling improvement. For example, a school value-added information system has been developed using test scores to discover quality schooling even when absolute scores are low. The Basic Competency Assessment (BCA) system gives teachers the ability to monitor and respond to student learning needs prior to Territory-wide School Assessment (TSA) evaluation testing. A third resource made available to schools enables monitoring of student affective and social outcomes (i.e., APASO) as an adjunct to achievement monitoring tools.

The reputation of schools, however, is often largely determined by absolute student scores, despite efforts to introduce value added information systems as a more defensible means of evaluating school quality. For example, the centrally administered TSA system has threatened many schools (Yu et al., 2006). Hong Kong teachers did not consider government assessment for learning changes as equally important as the need to prepare students for high-stake examinations (Chan, 2007).

The use of public examinations for selection of students, evaluation of schools and teachers is hardly less aggressive in China proper. Indeed, examinations remain an important part of assessment cultures in many Asian countries and their influence needs to be taken into account when assessment reforms are discussed (Kennedy, 2007). Hence, we should expect teachers in Chinese societies to have quite different conceptions of assessment to westerners.

1.4. Preliminary studies of Hong Kong teacher conceptions of assessment

After careful translation of the TCoA into Chinese, a survey of nearly 300 primary and secondary school teachers in Hong Kong was carried out (Brown, Kennedy, Fok, Chan, & Yu, 2009). The fit of the model was marginal and improved somewhat when mapped to a newly developed Assessment Practices Inventory. The most important different feature of the TCoA results was that among Hong Kong teachers, there was a strong and positive correlation ($r = .91$) between the conception that assessment evaluates students and assessment is for improvement. In New Zealand, the same two conceptions were very weakly correlated ($r = .21$). This difference was attributed to cultural features of the Confucian system in Hong Kong which emphasizes educational testing as a force for improved learning.

1.5. Preliminary studies of mainland China teacher conceptions of assessment

A small study of nearly 100 polytechnic lecturers in southern China surveyed their conceptions of assessment using the full 50 item version of the TCoA inventory (Li & Hui, 2007). The lecturers agreed most of all that assessment improves quality of teaching and that it makes schools and teachers accountable; they rejected the conception that assessment was bad or ignored. Interestingly, the two accountability conceptions tended to correlate with the assessment is valid and descriptive factors, leaving the two improved teaching and learning factors in a separate factor. It was argued that assessment was viewed this way because of competitive pressures to demonstrate to industry that the institute was delivering high-quality students for employment in the industry. In this way, lecturers made a distinction between evaluative and educationally functional purposes of assessment.

2. Creating a teachers' conceptions of assessment inventory for Chinese contexts

These preliminary studies with the TCoA in Chinese contexts suggested very clearly that the TCoA inventory tapped into just some of the important aspects of how Chinese teachers understand the use and purpose of assessment. Further, it was clear that the accountability conceptions were conceived of in quite a different manner to New Zealand and Queensland. Hence, a project was initiated to develop an inventory that captured teachers conceptions of assessment in Chinese contexts.
2.1. The joint HKIEd-SCNU research project

A collaborative research project was initiated in 2008 and completed a series of studies in Hong Kong and the Guangdong province of China. These two regions are contiguous and have populations that are overwhelmingly Han Chinese. However, there are significant differences between the two regions. Guangdong is fully part of the People’s Republic of China, uses Putonghua as the official medium of instruction, and provides only 9 years of compulsory schooling to its residents. In contrast, Hong Kong is a Special Administrative Region of China with considerable political, economic, and social autonomy, uses both Cantonese and English as the media of instruction, and provides 11 years of compulsory schooling to its residents. In terms of assessment policy, Hong Kong has adopted an assessment for learning policy while retaining high-stakes public examinations and school-based end-of-year examinations. Guangdong, on the other hand, has highly competitive entrance and exit testing systems to stream students into highly selective schools at the start of primary, middle, and secondary schooling. Thus, we could expect that between these two jurisdictions there may be strong similarities and distinctions in how teachers conceive of assessment.

2.2. Research questions

The goal of this study was to identify additional conceptions of assessment held by teachers working in Chinese contexts and, subsequently, validate a new questionnaire. It was assumed that the four constructs embedded in Brown’s TCoA were valid, though potentially two of which were under-represented. Further, the goal was to establish a common questionnaire and measurement model across the two samples of teachers from South China and Hong Kong. The research questions addressed in this study were:

1. Can additional conceptions of assessment be identified in the responses of Chinese teachers to a self-report inventory?
2. What model (i.e., number of factors and their inter-relationship) fits the responses of teachers from South China and Hong Kong?
3. To what degree is the model statistically equivalent for both groups?
4. To what degree are the factor mean scores equivalent for both groups?
5. Are differences in the model or mean scores consistent with jurisdictional differences between South China and Hong Kong?

2.3. Development of the TCoA instrument for Chinese contexts

In response to the previous studies described in 1.4 and 1.5, a series of small scale studies were implemented with a view towards identifying conceptions of assessment missing from the original research tool. An analysis of Hong Kong primary school curriculum leader (n=22) opinions about the uses and purposes of assessment identified three additional purposes that were associated with the notion of improvement (Hui, in press). These were to change students’ attitudes towards learning; identify their potential; and prepare students for future challenges: all of which are consistent with the Chinese idea that high assessment results indicate a better or more valuable person. A parallel series of interview studies conducted in China (Wang, 2010) identified the notions that assessments are used to prepare students for high-stakes and/or externally administered examinations or tests and to control students’ behavior both in- and out-of-class.

Consequently, a new questionnaire was developed around six controlling constructs (i.e., assessment makes schools accountable; assessment makes students accountable; assessment improves teaching and learning; assessment develops students into better people; assessment is used to control both students and teachers; and assessment is irrelevant) (definitions in Appendix A). Compared to the original TCoA, this framework has introduced two new constructs (Development and Control) and added new items and meanings to the existing four constructs (Student Accountability, School Accountability, Improvement, and Irrelevance).

The items for the Chinese–Teachers’ Conceptions of Assessment (C-TCoA) inventory were drafted simultaneously in three languages (i.e., English, Cantonese, and Putonghua) with the goal of achieving functional equivalence based on a decentered approach (Werner & Campbell, 1973). Modifications to items were made in each language to obtain natural and appropriate versions of the items in each language and still have the same meaning. In accordance with procedures outlined by Gable and Wolf (1993), small samples of teachers were asked to (1) classify each item according to the definitions for each of the six constructs and (2) evaluate the equivalence of the Cantonese and English versions. Where problematic items were identified, revisions were made to the construct definitions, items, or the wording of items in one or all languages.

2.4. Survey participants

Participants were recruited through purposive sampling since cold-calling, random sampling is a relatively ineffective means of recruitment in Chinese school contexts. This means that schools known to the researchers and their own graduate students were used to introduce the survey to teachers. Teachers were recruited in Hong Kong and Guangzhou cities, both large metropolitan areas in the Cantonese region of China. In Hong Kong, 1464 teachers were approached through personal referrals and 1014 inventories were returned, for a return rate of 69%. In Guangzhou, data were collected from five secondary
and four primary schools, with a return rate of approximately 80% (Gao et al., 2010). These robust response rates are a consequence of the strong working relationships developed between the school principals and research institutions and the generally high level of cooperation between teachers and school leaders in Chinese societies.

The samples from both jurisdictions were remarkably similar for sex and ethnicity characteristics (i.e., predominantly female and Chinese) (Table 1). In terms of teaching experience, level being taught and subjects taught, the samples were less comparable as a consequence of recruitment processes. Over half of the south Chinese teachers had less than 10 years teaching experience, while half of the Hong Kong teachers had more than 10 years experience. The Hong Kong sample was dominated by primary teachers, while the South China sample had a majority of secondary teachers. Over half of the Hong Kong teachers were either Chinese or English teachers, while only one-third of South China teachers had these languages as their teaching subject and nearly a quarter were science teachers. If statistically equivalent models are identified, despite these differences in teaching experience, subject, and level, it could be argued that the results are a function of shared Chinese cultural values.

2.5. Analyses

To answer our first research question, exploratory and confirmatory factor analyses were used to develop a well-fitting model that explained Chinese teacher responses to the C-TCoA. Confirmatory factor analysis was used to test the hypothesised model (Klem, 2000; Hoyle, 1995), exploratory factor analysis was used to develop an alternative model, and confirmatory approaches were used to validate the fit of the alternative trimmed model. Maximum likelihood estimation with oblique rotation was used in exploratory factor analysis (Costello & Osborne, 2005). A conventional approach was taken to determining the number of potential factors and their members: factors had to have (1) eigen-values > 1.00, (2) at least three items which were conceptually aligned, (3) items had pattern or regression loadings of > .30, and (4) all cross-loadings < .30 (Bandolos & Finney, 2010).

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 \( \chi^2 \) 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. All analyses were carried out in AMOS (Arbuckle, 2008) using Pearson product moment correlations. All cases with more than 10% missing responses were removed and remaining missing values were imputed using the expectation maximisation procedure (Little & Rubin, 2002), resulting in no missing data.

To test for equivalence of the model across the two samples, nested, multi-group invariance analysis (Byrne, Shavelson, & Muthen, 1989) was conducted. This involves constraining the model to be equivalent for each a parameter, examining the fit statistics, and moving to test the next parameter only if the fit criteria indicated that the parameter values were equivalent.
Testing stops when a parameter is shown not to be equivalent. 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 from factor to items and among factors (i.e., configural equivalence),
(2) all regression 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.

There is consensus that equivalence of item and/or factor residuals is not required to argue for equivalence. 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, & Jumbo, 2007). The equivalence of the pathways is accepted if the RMSEA for a multigroup model is ≤.05. Given the large sample sizes, it was decided to examine the change in CFI, rather than the difference in \( \chi^2 \), to determine whether equivalence was demonstrated. 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 ≤.01 (Cheung & Rensvold, 2002).

Given that the structure of conceptions exists in the two groups, we would expect that any effect of context would be manifest in mean score differences for the model factors. 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. To establish the practical significance of differences in factor mean scores, the difference in mean scores was calculated as Cohen's (1977) effect size \( (d) \). Hattie (2009) has shown that in education research absolute values of \( d \) up to .20 are trivial, between .21 and .39 are small, between .40 and .59 are moderate, and >.60 are large.

3. Results

3.1. Factor analysis

Exploratory factor analysis was embarked upon because the expected six-factor solution could not be recovered. Initially a nine-factor solution was detected, but because the content of two factors seemed to duplicate factors with stronger eigenvalues, a seven-factor solution with 31 items (Appendix B) was tested in confirmatory factor analysis. Factor 1 describes assessment for student betterment or development (see Section 1.3 for description); Factor 2 focuses on the irrelevance and negative aspects of assessment; Factor 3 identifies examinations as assessment; Factor 4 recommends that teachers take into account measurement error when using assessments; Factor 5 focuses on assessment to help students learn; Factor 6 shows that assessment is used to control teachers and evaluate schools; and Factor 7 indicates that assessments are reliable and accurate.

This seven factor inter-correlated solution was tested with all participants and found to have acceptable fit \( (\chi^2 = 3479.15, df = 414; \chi^2/df = 8.40, p < .001; CFI = .87; RMSEA = .062; SRMR = .057; \gamma = .91) \). However, close inspection of the factor inter-correlations and the conceptual meaning of the factors suggested that a simplifying second-order structure may be present. Hence, it was decided to test a hierarchical model with 3 intercorrelated conceptual factors.

- Metafactor 1 = Improvement containing F7 Accuracy, F5 Help Learning, & F1 Student Development
- Metafactor 2 = Accountability containing F4 Error, F3 Examinations, F6 Teacher and School control
- Metafactor 3 = Irrelevance factor alone

The hierarchical model had somewhat worse fit but still within standards for not being rejected \( (\chi^2 = 3856.97, df = 426; \chi^2/df = 9.05, p < .001; CFI = .85; RMSEA = .065; SRMR = .065; \gamma = .90) \). The advantage of this model was that it was simpler and drew attention to three core purposes of assessment. Instead of 7 inter-correlated factors, there were 3 multi-faceted and inter-correlated purposes in how teachers conceived of assessment.

3.2. Comparing Hong Kong and South China

The invariance of the hierarchical model was tested with two groups (i.e., Hong Kong and South China). The equivalence of regression weights from 1st-order factors to items was tested (Model 1). Assuming those differed within chance, the equivalence of 2nd-order factor regressions to 1st-order factors was tested (Model 2). Then the equivalence of the covariance matrix among the 3 metafactors was tested (Model 3). Then the equivalence of the metafactor residuals was tested for equivalence (Model 4). Finally, the residuals for each item were tested for equivalence (measurement residuals—Model 5). Statistical equivalence between the HK and SCNU samples for all parameters was found except for item residuals (Table 2).

The fit of the 2-group model constrained to all parameters except measurement residuals equivalent was (Model 4) good \( (k = 62; \chi^2 = 4612.89, df = 891; \chi^2/df = 5.18, p = .02; CFI = .83; RMSEA = .047; SRMR = .065; \gamma = .94) \). Hence, this analysis supports the interpretation that the constrained equivalent hierarchical model does not have to be rejected and that the two groups of teachers responded to the inventory in equivalent fashion, despite differences between jurisdictions (Wu et al.,
Table 2
Nested progressively constrained models 2-group confirmatory factor analysis fit statistics.

<table>
<thead>
<tr>
<th>Model</th>
<th>CFI</th>
<th>ΔCFI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 0. Unconstrained</td>
<td>.840</td>
<td>-</td>
</tr>
<tr>
<td>Model 1. Equivalent item regressions</td>
<td>.839</td>
<td>.001</td>
</tr>
<tr>
<td>Model 2. Equivalent factor regressions</td>
<td>.839</td>
<td>.000</td>
</tr>
<tr>
<td>Model 3. Equivalent factor covariances</td>
<td>.836</td>
<td>.003</td>
</tr>
<tr>
<td>Model 4. Equivalent Factor residuals</td>
<td>.834</td>
<td>.002</td>
</tr>
<tr>
<td>Model 5. Equivalent item residuals</td>
<td>.775</td>
<td>.059</td>
</tr>
</tbody>
</table>

Notes: CFI, comparative fit index; ΔCFI values < .01 indicate invariance.

2007). Hence, any differences in mean scores can be attributed to differences in populations, rather than to differences in responding. Note that the constrained Model 4 means that the values for the covariance matrix among the three metafactors and the structural regression weights are identical for both groups.

Factor loadings for items for the constrained Model 4 are shown in Table 3. Structural regression weights from Improvement to the three contributing 1st-order factors were strong (i.e., Accuracy $\beta = .92$, Student Development $\beta = .77$, and Help Learning $\beta = .75$). The structural regression weights from Accountability to two of the three contributing 1st-order factors were strong (i.e., Examination $\beta = .84$, Teacher and School Control $\beta = .83$, and Error $\beta = .35$). The intercorrelation between Accountability and Improvement was strongly positive ($r = .80$), while the intercorrelations with Irrelevance were weak (i.e., Improvement–Irrelevance $r = -.22$; Accountability–Irrelevance $r = .28$).

Having established a statistically invariant model of the C-TCoA for both samples, it was possible to examine differences in mean scores. Table 4 provides descriptive statistics for each factor and metafactor for the China and Hong Kong samples and Cohen’s $d$ statistic for the differences in mean scores. Consistent with the overall fit statistics from the confirmatory factor analysis, the Cronbach’s alpha estimates of scale reliability were all acceptable to good. The mean scores fell in a somewhat narrower range among the China teachers compared to the Hong Kong sample. The effect sizes ranged from trivial to large, suggesting jurisdiction related differences seemed to exist for some of the C-TCoA scales.

Table 3
Teacher conceptions of assessment in Chinese contexts confirmatory factor analysis standardized factor loading for joint Hong Kong and South China solution.

<table>
<thead>
<tr>
<th>Item</th>
<th>Standardized factor loadings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Student development</td>
</tr>
<tr>
<td>22</td>
<td>.79</td>
</tr>
<tr>
<td>20</td>
<td>.77</td>
</tr>
<tr>
<td>24</td>
<td>.77</td>
</tr>
<tr>
<td>14</td>
<td>.71</td>
</tr>
<tr>
<td>39</td>
<td>.76</td>
</tr>
<tr>
<td>62</td>
<td>.71</td>
</tr>
<tr>
<td>40</td>
<td>.68</td>
</tr>
<tr>
<td>33</td>
<td>.67</td>
</tr>
<tr>
<td>38</td>
<td>.62</td>
</tr>
<tr>
<td>44</td>
<td>.62</td>
</tr>
<tr>
<td>25</td>
<td>.58</td>
</tr>
<tr>
<td>7</td>
<td>.54</td>
</tr>
<tr>
<td>36</td>
<td></td>
</tr>
<tr>
<td>58</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td></td>
</tr>
<tr>
<td>61</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td></td>
</tr>
</tbody>
</table>
The Hong Kong teachers agreed more with all facets of the Improvement metafactor and considerably more with the Help Learning and Accuracy factors. The HK teachers were moderately stronger on the examinations component of accountability. The China teachers were moderately stronger on the irrelevance factor.

4. Discussion

4.1. Summary of results

A hierarchical, inter-correlated factor structure fitted well to the responses given by two large samples of teachers from Hong Kong and the South China province of Guangdong. The model has identified seven factors, six of which aggregate into two second-order factors that inter-correlate with each other and the seventh factor of irrelevance. Four of the factors were consistent with the hypothesised constructs of Developmental, Irrelevance, Improvement, School Accountability, and Control. In contrast, three much more narrow and technical factors were identified (i.e., examination, error, and accuracy). The model has successfully identified additional constructs (i.e., examination, control, development) that had not been part of the earlier TCoA work with western teachers.

Where the mean score differences were more than trivial, the Hong Kong teachers had a higher mean except for the Irrelevance factor which was endorsed more strongly by the China teachers. If there had been a response bias, a higher mean from Hong Kong teachers would have been expected across all factors, which is not apparent. Hence, we can suggest that the mean scores are indicative of real world differences in teacher conceptions of assessment. While there are striking and intriguing similarities with western teachers studied with the TCoA, the current results support the adoption of a revised and extended Chinese-Teachers’ Conceptions of Assessment inventory for use in Chinese contexts. However, this model makes several telling changes.

4.2. Understanding assessment as accountability

Accountability integrates the earlier distinction between evaluating students and schools and subsumes those as part of control. Accountability in the Chinese context is about controlling schools, teachers, and students; not simply determining how good they are. Furthermore, accountability incorporates both assessment as examinations and teachers’ taking into account measurement error. In earlier studies with New Zealand and Queensland teachers (Brown, 2008), Error had been part of Irrelevance rather than Accountability. The current association seems to indicate that schools, students, and teachers are controlled by examinations but teachers ought to take into account their inherent error. This would imply teachers are sensitive to negative consequences due to faulty interpretations of examination performance that have large consequences for participants (e.g., failure, graduation, selection, public praise or condemnation). It would seem teachers believe there is a need to balance pass-fail or other categorical decisions by the margin of error in every examination score.

Nonetheless, consistent with earlier studies with Chinese teachers (Brown et al., 2009; Li & Hui, 2007), accountability was positively correlated with improvement (r = .80). This association indicates that, insofar as these two jurisdictions are concerned, teachers are persuaded that a powerful way to improve student learning is to examine them. Since the Chinese tradition of examination-merit decisions is so long-standing and because it is so powerful in contemporary China and Hong Kong, it seems highly reasonable to believe that accountability assessments function to improve teaching and learning. A similar association, albeit much weaker, has only been seen among secondary teachers in Queensland and New Zealand (Brown, 2008). This suggests that to the extent that assessment for learning policies advocate no testing or examinations, there will difficulties in their adoption by teachers in Chinese contexts.
4.3. Understanding assessment for improvement

The Improvement factor invoked helping learning, confidence in the reliability of assessments, and introduced a more complex, richer construal of development than was previously detected. This last result is consistent with Hui (in press) in which experienced primary school curriculum leaders saw assessment as helping to make students better people. Hence, improvement is a relatively unproblematic construct—assessment leads to improving student learning and personal development; provided that it is accurate. This emphasis that assessment contributes to holistic development may appear alien to the western tradition, where clear separation of academic and affective components in reporting school performance is encouraged (Friedman & Frisbie, 1995). Yet in Chinese contexts, high achieving students are accorded higher status both from their peers and their families. Thus, it would seem such high stakes consequences have considerable potential to motivate students to learn.

4.4. Understanding the irrelevance of assessment

This model identifies irrelevance as a real factor with quite an independent character. Unlike previous studies where irrelevance was strongly inverse to improvement, this study shows that it is only weakly inverse. Hence, it seems that teachers had the opinion that while assessment is intended for improvement, it is only weakly relevant to real improvement. Nonetheless, the pattern of inverse correlation to improvement and positive correlation to accountability, a statistically significant difference, is important. Assessment for accountability invokes a sense of irrelevance and rejection, while assessment for improvement invokes the opposite. In this way, the Chinese teachers are very similar to previously studied western teachers. This suggests that the validity of the accountability system was being questioned. Given the reluctance of Chinese people to give negative evaluations (Bond & Hwang, 1986), especially of formal authorities (McCormick & Shi, 1999), this result may be an underestimate of how strongly these teachers question the validity of the accountability-examination purposes.

4.5. Understanding the role of error in assessment

The model suggests that taking error into account is a response associated with the use of examinations for school and teacher accountability. In contrast, assessments used for improvement are bundled with the notion that assessments are reliable and accurate. This generates an interesting insight into teachers’ concerns about assessment usage. It would appear that teachers believe evaluative judgements about teachers and schools need to be adjusted by the measurement error of the examinations, while judgements about improvement depend on assessments that are accurate and reliable. This result is consistent with recent UK research (He, Opposs, & Boyle, 2010) that has shown that teachers are generally concerned about and intolerant of error in the high-stakes examinations that impact upon student life chances and perhaps school and teacher reputation. This suggests Chinese teacher confidence in lower-stakes, standardized tests would likely to be very high, provided, of course, that tests can be demonstrated to be both valid and accurate, a somewhat daunting challenge but one which appears to be somewhat addressed in Hong Kong (Mok, 2007). Additionally, teacher confidence in the accuracy of assessments may arise from confidence in their own professional judgements of student performance as much as it is from quality assessment tools. Thus, additional research is needed to understand better what kinds of assessment practices are deemed to be improvement-oriented and reliable; do teachers mean formal tests, their own overall judgements of student work, or some combination? Hence, development and support of teacher use of both formal and professional-judgement resources may help teachers deliver improvements in student learning.

4.6. Contextual differences across Chinese contexts

It would appear, based on the conceptions of Hong Kong teachers, that improvement-oriented assessments (i.e., formal and informal mechanisms) are considered accurate and viewed relatively positively. This appears consistent with the presence of an improvement-oriented assessment policy in Hong Kong and the availability of low-stakes, high-quality assessments (e.g., such as the BCA and APASO) and a school-based quality assurance system (Mok, 2007). This positive view of assessment appears to spill over to the public accountability system in Hong Kong which is given weak but positive endorsement. The general attitude of Hong Kong teachers was relatively not negative towards assessment in general. In contrast, the Chinese teachers endorsed the irrelevance of assessment more and gave less support to its improvement and accountability orientations. It may be that the high-pressure selection-orientation of examinations is much greater in China and that there is less confidence in the array of assessment resources available to support improved learning outcomes for students and in the methods use to evaluate teachers and schools.

4.7. Implications for future research

The strong similarity of the model for teachers from two parallel but distinct jurisdictions seems to suggest that there are parallel constructs associated either with Chinese identity or with a high-stakes, public examination controlled system. If the latter explanation is correct, then we should expect, in societies with high-stakes public examination cultures (e.g., Africa,
Latin America, and Asia), a very similar pattern of conceptions. We should expect wherever there is a strong perception of cheating, corruption, or favouritism in the implementation of the exam system that teachers will have strongly negative views about assessment.

Research in other Chinese contexts would go some way to determining the impact of different contextual factors (e.g., resourcing, medium of instruction, educational history, etc.) on variations within Chinese teacher conceptions of assessment. For example, Singapore, even more than Hong Kong, uses English as the medium of instruction, whereas Taiwan and the People’s Republic use only Mandarin. Both Hong Kong and Singapore share histories of British colonial governance and this interaction with western approaches to schooling may contribute to differing functions and conceptions of assessment relative to other Chinese jurisdictions. Furthermore, the large differences in annual per capita income between the People’s Republic (approximately $8000US) and the other Chinese jurisdictions (average about $60,000US) may contribute to both different schooling resources and, consequently, beliefs about educational practice. Given popular stereotypes about differences between northern and southern Chinese cultures (especially in terms of food, physical appearance, and accent) (Young, 1988), it seems appropriate to survey teachers in northern regions of China so as to ascertain whether such differences extend to educational belief systems.

Additionally, research with western teachers who have been strongly influenced by child-centred developmental agendas may find the new factor of student development a better way to capture how assessment is used and understood. Though, whether this construct would be accepted by western teachers as a legitimate function of assessment is an open question.

These insights raise the larger problem of the origins and plasticity of teacher beliefs concerning educational assessment. Pajares (1992) has pointed to the experiences teachers have as pupils and students as the origin of teacher beliefs, suggesting socialisation as the mechanism by which teacher beliefs about education are formed. Students absorb beliefs about assessment from their parents and teachers and, as teachers, pass them on to future generations of students. It should also be remembered that the more implicit and experiential a belief is the less likely it can be changed easily (Brown, 2008). While teachers are meant to be professional agents capable of making changes within educational processes, strong accountability pressures from the system and society in general are likely to reduce the possibility of belief and practice changes. Nonetheless, if changes are made by the relevant accountability agency (e.g., Hong Kong Education Bureau) as to how assessment is used (e.g., remove high-stakes examinations as the mechanism by which progress in the system is evaluated) teachers’ beliefs about assessment may become coherent with the new rationality of the environment.

An interesting study, yet to be conducted, to examine the sensitivity of teachers’ assessment conceptions to contextual factors would be to experimentally require responses, not to assessment in general as in this study, but rather to specific types and forms of assessment (e.g., high-stakes examinations vs. low-stakes diagnostic evaluations). If responses are statistically different dependent on the type of assessment, there would be further evidence that beliefs about assessment depend on contexts. While there are studies that show teachers can implement improvement-oriented assessment practices within high-stakes examination contexts (e.g., Black, Harrison, Lee, Marshall, & Willian, 2003 in the UK; Carless, 2011 in HK), we do not yet know if those changes came about because teachers changed their beliefs about assessment or were enabled to enact their pre-existing beliefs within an environment generally inimical to those beliefs. Use of the C-TCoA, as part of the studies described here, may help resolve how plastic or inflexible teacher conceptions of assessment are.

4.8. Policy Implications

The current study is a step towards understanding teacher conceptions of assessment in Chinese contexts and towards developing an understanding of how teacher beliefs both reflect and construct societal understandings that legitimate how assessment is used in an educational system. Education and qualifications policy developers should be concerned by the positive association of accountability mechanisms with irrelevance, while being reassured by the negative connection between improvement and irrelevance. The perception that accuracy is associated with improvement while inaccuracy is connected with accountability should also give pause to those responsible for systems. Clearly, work needs to be done, not just to persuade teachers that their concerns are ill-founded, but, more importantly, to provide high-quality assessment tools and methods that permit improvement and which reduce consequences associated with high-stakes decisions within margins of error (e.g., allowing all students within the margin of error of a cut-score to be given the grade as if they had exceeded the cut-score).

The results from the Hong Kong sample seem to indicate that policy, resource, and professional development initiatives towards an assessment for learning paradigm are supporting somewhat more positive views of assessment; this may provide a model for other Chinese contexts. Nonetheless, consistent with Kennedy et al. (2008), we believe the strong positive association of accountability and improvement means that a full-blown implementation of a western assessment for learning system (e.g., Queensland’s no examination system—Cumming & Maxwell, 2004) is unlikely to be successful in Chinese contexts. The overwhelming message of this study is that Chinese teachers conceive of accountability as highly unified with educational improvement.

Additionally, we might speculate that changes in assessment policy, practice, and resources that do not reflect the dominant social construction of assessment will have little chance of success in changing teachers’ thinking about the nature and purpose of assessment. Whether assessment reform based on individualist views of learning and assessment is more difficult in collectivist societies, such as those in this study, is an interesting cross-cultural question. The C-TCoA inventory
used in this study can make a contribution to continuing research in exploring the cultural construction of classroom assessment.

Acknowledgements

A previous version of this paper was presented at the 2010 biannual International Test Commission conference in Hong Kong. Funding for this study was provided by the Hong Kong Institute of Education Blueprint Grant #H056 to the Teachers’ Conceptions and Practices of Assessment in Chinese Contexts project in collaboration with South China Normal University, Guangzhou, China. Professor Gao Lingbiao and his team at SCNU are thanked for their contribution.

Appendix A. Construct definitions

School accountability: Assessment holds teachers, schools, and systems accountable for achieving societal goals and expectations. This is usually done using student performance on external, high-stakes examinations or tests. Assessment results are used to demonstrate publicly that teachers and schools are doing a good job. Schools and teachers are rewarded (e.g., pay bonuses) or punished (e.g., dismissal) for exceeding or not reaching required standards.

Student accountability: Assessment holds students accountable for learning what was expected of them by society. This is usually done using performance on examinations or tests. This requires grading, scoring, or evaluating student performance against standards, objectives, targets, or expectations. Students experience positive or negative consequences (e.g., placement into classes or groups, selection for special programs, or awarding of certificates) depending on their performance.

Improvement: Assessment is a means of improving the quality of both students’ learning and teachers’ instruction. A variety of assessment techniques are used to identify the content and processes of student learning, as well as the quality of instruction. The goal is answering accurately two key questions: “who has learned what” and “who needs to be taught what next”.

Developmental: Assessment cultivates positive moral and ethical qualities and values in students which contribute to their lifelong and life-wide learning and good citizenship. A wide variety of valued personal and social skills appropriate to full participation in society are developed. The goal is to help students develop positive social conduct, moral character, and appropriate personal potential and qualities.

Control: Assessment controls student behavior and actions both in and out of class. Assessment is used managerially to control schools or classrooms. The assessments are not necessarily scored or recorded, rather they lead to better discipline. Assessment is used to enhance and maintain the control of the teacher and the dominance of teacher’s opinions over those of the student.

Irrelevance: Assessment serves no legitimate role within teaching and learning. While assessments may be administratively required, teachers’ knowledge of students based on long relationship and their understanding of curriculum and pedagogy precludes the need for assessment. Externally mandated assessments have negative effects on teacher autonomy and professionalism, and distract from the real purpose of teaching (i.e., student learning). Since accurate and precisely correct measurement of assessment is difficult, teachers may have legitimate grounds to ignore assessment.

Appendix B. C-TCoA items by factor organised by metafactor

Metafactor, factors, and items

Improvement

Student development

9. Assessment helps students succeed in authentic/real-world experiences.

評估幫助學生獲取真實世界情境的經驗。

17. Assessment fosters students’ character.

評估培養學生個性。

10. Assessment is used to provoke students to be interested in learning.

評估用來激發學生的學習興趣。

13. Assessment stimulates students to think.

評估激勵學生思考。

11. Assessment cultivates students’ positive attitudes towards life.

評估培養學生正面的人生觀。

Help learning

1. Assessment helps students improve their learning.

評估有助學生改善學習。
2. Assessment determines if students meet qualification standards.
評估確定學生是否達標。

3. Assessment information modifies ongoing teaching of students.
評估的資料有助於不斷改進教學。

Accuracy

29. Assessment results are trustworthy.
評估結果是可靠的。

8. Assessment results can be depended on.
評估結果是可以信賴的。

4. Assessment results are sufficiently accurate.
評估結果是比較準確的。

Irrelevance

12. Assessment results are filed & ignored.
評估結果會被存檔而後置之不理。

18. Assessment interferes with teaching.
評估干擾教學。

15. Assessment is an imprecise process.
評估是一個不精確的過程。

7. Assessment has little impact on teaching.
評估對教學的影響微不足道。

27 Assessment forces teachers to teach in a way against their beliefs.
評估迫使教師用有違自己信念的方法教學。

Accountability

Examinations

23. Assessment helps students gain good scores in examinations.
評估讓學生在考試中取得好成績。

31. Assessment familiarizes students with examination formats.
評估讓學生熟悉考試模式。

19. Assessment teaches examination-taking techniques.
評估用來教授考試技巧。

22. Assessment sets the schedule or timetable for classes.
評估主導課堂教學的進度。

5. Assessment prepares students for examinations.
評估讓學生為應付考試作準備。

26. Assessment helps students avoid failures on examinations.
評估讓學生避免考試失利。

14. Assessment is assigning a grade or level to student work.
評估是為學生的學業評分或評級。

24. Assessment selects students for future education or employment opportunities.
評估是為未來升學或就業來挑選學生。

Error

21. Teachers should take into account error and imprecision in all assessment.
教師應該考慮評估的誤差和不精確性。

28. Assessment results should be treated cautiously because of measurement error.
評估結果應審慎運用，因量度有誤差。

Teacher & School Control

25. Assessment results contribute to teachers’ appraisals.
對學生進行評估所得到的結果有助於評價教師的表現。
20. Assessment indicates how good a teacher is.  
21. Assessment is an accurate indicator of a school's quality.  
22. Assessment measures the worth or quality of schools.  
23. Assessment is used by school leaders to police what teachers do.  

References  
Chan, J. K. S. (2007, May). We have various forms of assessments but only summative assessments count: Case studies of the implementation of an innovative assessment policy in Hong Kong. Paper presented at the Redesigning Pedagogy: Culture, Knowledge & Understanding Conference.  
China Civilisation Centre. (2007). China: Five thousand years of history and civilization. Hong Kong: City University of Hong Kong Press.  
Hui, S. K. F. Missing conceptions of assessment: Qualitative studies with Hong Kong curriculum leaders. The Asia-Pacific Education Researcher, in press.  


