Teachers' Conceptions of Assessment: Developing a model for teachers in Hong Kong

Gavin T L Brown
Sammy K F Hui
Flora W M Yu
The Hong Kong Institute of Education

Paper presented at the biannual conference of the International Test Commission, July 19-21, 2010, Hong Kong.

Correspondence concerning this article should be addressed to Dr Gavin T L Brown, Dept. of Psychological Studies, The Hong Kong Institute of Education, 10 Lo Ping Road, Tai Po, NT, HONG KONG SAR or by email to: gtlbrown@ied.edu.hk.

Abstract

Hong Kong has an assessment *for* learning policy and a cultural context that emphasizes examinations. In addition to associating student grading with improvement, important improvement-oriented conceptions have been identified among Hong Kong teachers and which were not fully instantiated in the original *Teachers' Conceptions of Assessment* (TCoA) inventory. An expanded Chinese-TCoA inventory was administered to in Chinese in both Hong Kong and southern China. The intended 6 construct structure was not supported. Exploratory factor analysis (MLE, oblimin rotation) identified 7 factors, which were further reduced to 3 major inter-correlated constructs (i.e., improvement, accountability, and irrelevance). Improvement and accountability were strongly inter-correlated (r=.80), while improvement had a weak negative correlation with irrelevance and accountability had a weak positive correlation with irrelevance. The model had 7 factors based on 33 items and acceptable fit (χ^2=3856.97, df=426; χ^2/df=9.05, p<.001; CFI=.85; RMSEA=.065; SRMR=.065; gamma=.90). Two group nested invariance testing (i.e., Hong Kong vs. southern China) showed that the model was statistically equivalent, except for item residuals (k=62; χ^2=4612.89, df=891; χ^2/df=5.18, p=.02; CFI=.83; RMSEA=.047; SRMR=.065; gamma=.94). Differences in mean scores between the two groups showed that Hong Kong teachers agreed more with the ideas of assessments helping student learning, being accurate, and being examinations; whereas southern China teachers agreed more that assessment was irrelevant. This study contributes to our understanding of how assessment is understood by teachers working within Chinese contexts.

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

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. 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.

It is expected that differences in culture or society lead not only to differing policies, but also distinctive conceptions of practices or processes. For example, Hamilton, et al. (2007) reported that teachers in California, Georgia, and Pennsylvania had very similar responses, experiences, and attitudes towards standards-based accountability assessments; they attributed this to similarities between the systems. Similarly, teachers in New Zealand and Queensland had very similar conceptions of assessment (Brown & Lake, 2006). Furthermore, Brown and Harris (2009) reported that recent, proactive school policies to use testing as part of school-wide improvement initiatives modified the beliefs teachers had as to the purpose of assessment; the improvement-oriented purpose was replaced by a school accountability purpose as the dominant conception. Hence, it would appear that how teachers understand and value the competing purposes of assessment is sensitive to both general and specific policy priorities.

The model underlying the research into teacher conceptions of assessment has twin, interacting tracks leading to student outcomes. In the words of Brown and Harris (2009, p. 70):

> The conceptions of both teachers and students are influenced by various policy directions and family priorities and these beliefs, in turn, guide their separate teaching and learning practices. These two pathways are shaped by and respond to societal and cultural contexts, meaning that there will be different beliefs and practices in differing social, ethnic, and cultural groups. Note that this model does not attempt to portray the complex paths leading to teachers' and students' conceptions, which have been hinted at in Pajares (1992). … teacher beliefs are seen as mediating between policy and outcomes, rather than as external to the implementation processes. Second, policy directions are seen as a function of priorities within society and culture, suggesting that variation in conceptions and practices within societal contexts will be less than

those between contexts. Third, students themselves are thought to have a strong contributing role in shaping their outcomes.

This means, assuming the proposition that societally-derived policy and cultural priorities shape teacher conceptions is valid, that in other societies with different frameworks, teachers' conceptions of assessment should not fit so readily the current model. Lack of fit would also suggest that the four main factors of the TCoA inventory may not be sufficient for use in other societies, even with high-quality translation/adaptation. Further, even if the same factors are apparent, it may be legitimate to expect the pattern and strength of the paths among the factors would be statistically not invariant for samples taken from quite different populations. In other words, while some factors may be stable across populations, we can legitimately expect the correlations between those factors to differ across societies.

The TCoA 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.

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). Most students focus their learning on what they think they will be tested: the test becomes the curriculum (Biggs, 1996). The reputation of schools is often largely determined by absolute student scores, despite efforts of the Education Bureau to introduce value added information systems into place as a more defensible means of evaluating school quality. For example, the centrally administered territory-wide assessment system has threatened many schools (Yu et al., 2006). Hong Kong teachers do not consider the recent changes in government assessment policies and practices towards more of assessment *for* learning 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 perspectives on assessment to westerners.

Preliminary Hong Kong TCoA studies.

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 et al., 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. However, the conception of assessment is for improvement was a negative predictor of using practices related to assessment for school accountability. This was considered a parallel result to a New Zealand study with primary school teachers (Brown, 2009) which found that the conception of assessment for school accountability predicted the use of assessments of deep learning. Perhaps there is a shared concern among school teachers across the two societies that school accountability pressures are somehow not well connected with improved learning outcomes.

More recently, Hui (2009) summarized qualitative analyses of primary school teacher and curriculum leaders' opinions of assessment. He reported that three additional purposes of assessment which did not appear in the TCoA. Specifically, assessment changes students' learning attitudes, assessment identifies student potentials, and assessments helps prepare students for future challenges. It was argued that these three conceptions of how assessment is used arise from current policy emphases in Hong Kong on developing children for life-long learning in a knowledge-economy of the 21st century. He also suggested that these conceptions are likely to be strongly associated with an overall emphasis on assessment for improvement.

Preliminary China TCoA Study.

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. While the latter result is consistent with the New Zealand studies, the higher level of agreement for the school accountability purpose is quite different. 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 it was claimed lecturers made a distinction between evaluative and educationally functional purposes of assessment.

These preliminary studies with the TCoA in Chinese contexts suggest very clearly that the current TCoA inventory taps into just some of the important aspects of how Chinese teachers understand the use and purpose of assessment. However, the current studies show clearly that the accountability conceptions are conceived of in quite a different manner to New Zealand and Queensland. Student accountability is seen as a form of improvement while school accountability may have some legitimacy through public access to examination results. There are some clues in both TCoA survey studies to indicate that teachers make a distinction between improved learning outcomes and school evaluation or accountability. Whether their views would become more like those of teachers in low-stakes environments after the introduction of a policy that reduces consequences to schools from public examinations would be one way to determine whether these differences are attributable to culture or government policy.

The joint HKIEd-SCNU research project

A collaborative research project into teachers' conceptions of assessment in Chinese contexts was initiated in 2008 and has 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.

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. Most importantly, the goal was to establish a factor structure that was consistent with previous studies in which strong positive relationships could be seen between improvement and accountability conceptions. Hence, the research questions addressed in this study were:

1. Can additional constructs suggested as valid in Chinese contexts be identified in the responses of Chinese teachers to a self-report conceptions of assessment 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?

Development of a Chinese contexts TCoA instrument

In responding to the results reported in Brown et al. (2009), 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, 2009). These were to change students' attitudes towards learning, identify their potential, and prepare students for future challenges. A parallel series of interview studies conducted in China (Wang, 2009) identified the notions that assessments are used to prepare students for high-stakes and/or externally administered examinations or tests and to control students' behaviour 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 were drafted simultaneously in three languages (i.e., English, Cantonese, and Putonghua) with the goal of achieving functional equivalence. Hence, a decentered approach

(Werner & Campbell, 1973) to the drafting was taken and modifications were made in each version by the project team to obtain natural and appropriate versions of the items in each language and which had 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 both languages.

Analyses

To answer our first research question we used exploratory and confirmatory factor analyses to test a factor structure of six constructs and 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; Thompson, 2000), 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 eigen-values>1.00, at least three items which were conceptually aligned, all of which had pattern or regression loadings of >.30 and all cross-loadings <.30 (Bandalos & 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 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. 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 estimated using the expectation maximisation procedure (Little & Rubin, 2002) and so all analyses were carried out with 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.

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). However, 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, & Zumbo, 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 χ^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; Wu, Li, & Zumbo, 2007).

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 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 values of *d* up to .20 are trivial, between .21 and .39 are small, between .40 and .59 are moderate, and >.60 are large.

Results

Factor analysis

Exploratory factor analysis resulted in a nine-factor solution. However, inspection of the content in Chinese suggested that two factors were conceptual duplicates of previous factors. Hence, a 7 factor solution (Table 1) with 31 items (Appendix B) was tested in confirmatory factor analysis. Factor 1 describes holistic student development; 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.

Insert table 1 about here

This seven factor inter-correlated solution was tested with all participants and found to have acceptable fit (χ^2=3479.15, *df*=414; χ^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, exploratory factor analysis of the 7 factor scores was undertaken. Two factors had eigen values greater than 1.00. The first factor had 5 scales, the second factor was factor 2 by itself, and Factor 4 Error had weak loadings (<.30) on both factors. This result clearly suggested two dominant superordinate factors (i.e., improvement with exams and irrelevance), with F4 Error appearing to be independent. Hence, it was decided to test a hierarchical model with 3 intercorrelated major factors.

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

The unconstrained model (Figure 1) had somewhat worse fit but still within standards for not being rejected (χ^2=3856.97, *df*=426; χ^2/df=9.05, *p*<.001; CFI=.85; RMSEA=.065; SRMR=.065; gamma=.90).

Insert Figure 1 about here

The advantage of this model, however, is that it is much simpler to explain and draws attention to three core conceptions of assessment. Instead of 7 factors, there are 3 multi-faceted

conceptions which are inter-correlated. Note that each meta-factor is a strong predictor of its subordinate factors, with the exception of F4 Error which has a regression of only .35. However, this model provides the strongest and most conceptually meaningful result of all other options tried.

Comparing Hong Kong and South China

The invariance of the hierarchical model was tested with two groups (i.e., Hong Kong and South China). This means first 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 meta-factors 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; χ^2=4612.89, df=891; χ^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 as if they were members of the same population (in this case Chinese teachers). Further, any differences in mean scores can be attributed to differences in samples, rather than differences in responding. Note that the constrained Model 4 means that the values for the covariance matrix (i.e., inter-correlations between the thee metafactors) and the structural regression weights are identical for both groups.

Insert table 2 about here

Having established a statistically invariant model of the C-TCoA for both samples, it was possible to examine differences in mean scores. Table 3 provides descriptive statistics for each factor and meta-factor for the China and Hong Kong samples and Cohen's *d* statistic to show the scale of differences in mean scores. The Cronbach's alpha estimates of scale reliability are all greater than .60 with a mean of .78 for the China sample and .79 for the Hong Kong sample. These values are consistent with the overall fit statistics from the confirmatory factor analysis. The mean scores fell in a somewhat restricted range of 2.65 to 4.23 for the China sample and 2.28 to 4.65 for the Hong Kong sample. The effect sizes ranged from trivial (d<.20) (i.e., F1, F4, and F6) to large (d>.60) (i.e., F5, F7). This suggests that there are jurisdiction related differences for some of the C-TCoA scales.

Insert table 3 about here

The Hong Kong teachers agreed more with all facets of the Improvement meta-factor 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. When combined with the factor inter-correlations an interesting story emerges.

Discussion

A hierarchical, inter-correlated factor structure has been found that fits 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. It is noteworthy that four of the factors in this list are consistent with the hypothesised constructs 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 work with western teachers. This conceptual framework is structurally similar to the earlier model of the abridged TCoA developed in New Zealand and Queensland which also was inter-correlated and had two hierarchical factors.

However, this model makes several telling changes. Accountability integrates the earlier distinction between evaluating students and schools and subsumes those as part of control; a much more powerful notion than simply evaluating. 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 was part of Irrelevance rather than accountability. This 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 generated around cut-score boundaries—interpretations of examination performance that have large consequences for participants (e.g., failure, graduation, selection, public praise or condemnation) need to be balanced by the margin of error in every examination score. What is missing in this study is a greater sense of how teachers in Chinese contexts respond to the negative aspects of high-stakes examinations which control schooling and which may not be interpreted properly.

Nonetheless, consistent with earlier studies with Chinese teachers (Brown et al., 2009; Li & Hui, 2007), accountability is 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 examinations for accountability function to improve both 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 teachers have adopted child-centred, no-testing pedagogies (as encouraged by assessment for learning policies) will have a great deal of difficulty with the public examination systems implemented in Hong Kong and China.

The Improvement factor invokes helping learning (as expected from previous studies), confidence in the reliability of assessments (as expected from previous studies with the TCoA), and introduces a more complex, richer construal of development than was previously detected. This last result is consistent with the qualitative study reported by Hui (2009) where 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 appears alien to the western tradition, where clear separation of academic and affective components in reporting school performance is encouraged (Friedman & Frisbie, 1995). Whether assessments can validly make students better people, it seems certain that the high-stakes consequences motivate students to work harder. This in itself may be construed as making students better people. Although, whether this would be seen as enough is open to question.

This model identifies irrelevance as a real factor which has quite an independent existence to these teachers. Unlike previous studies where irrelevance was inversely and

moderately correlated with improvement, this study shows that it is only weakly inverse to improvement. Hence, it seems that teacher responses express the opinion that while it is intended for improvement, it may well still be irrelevant. Nonetheless, the pattern of inverse correlation to improvement and positive correlation to accountability, which reflects 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 is being questioned; given the positive response style of Chinese people to rarely give negative evaluations, this result is probably an underestimate of how strongly these teachers question the validity of the accountability-examination systems.

The strong similarity of the model in terms of how teachers from two parallel but distinct jurisdictions responded to the inventory seems to suggest that there may be 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 strong public examination systems (e.g., Africa, Latin America, Asia) a very similar pattern of results. Research in other Chinese contexts would go some way to determining the impact of different language and policy priorities on teacher conceptions of assessment. For example, Singapore uses English almost exclusively, whereas Taiwan uses only Putonghua. Furthermore, teachers in northern, western, and eastern China do not share the Cantonese language of Guangdong and Hong Kong and this may change responding to the inventory. 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.

The model suggests that among all teachers in this survey taking error into account is a response to the use of examinations for school and teacher accountability. In contrast, assessments used for improvement are considered reliable and accurate. This generates an interesting insight into teachers' concerns about assessment usage. Judgements about teachers and schools need to be adjusted by the measurement error of the tools, while judgements about improvement depend on assessments that are accurate and reliable. This suggests teacher confidence in lower-stakes, standardized tests should be very high—a view advocated in New Zealand where the government has supplied school-controlled, standardized testing systems (Hattie & Brown, 2008; Hattie, Brown, & Keegan, 2003). Hence, development and support of teacher use of such resources may further enhance teacher ability to deliver improvements in student learning.

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 conclude that the teachers are indicating a real world difference in their conception of assessment.

It would appear, based on the conceptions of Hong Kong teachers, that improvement-oriented assessments are considered accurate and seen relatively positively. This appears consistent with the presence of in Hong Kong of low-stakes, high-quality assessments such as the BCA and APASO and a strong improvement-oriented assessment and educational policy issuing from the EDB. This positive view of assessments appears to spill over to the public examinations system in Hong Kong which is given weak but positive endorsement. The general

attitude of Hong Kong teachers is relatively **not** negative towards assessment in general. In contrast, the China teachers endorse the irrelevance of assessment more and give 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.

While there are striking differences 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 for use in Chinese contexts.

References

Bandalos, D. L., & Finney, S. J. (2010). Factor analysis: Exploratory and confirmatory. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 93-114). New York: Routledge.

Biggs, J. (1996). The assessment scene in Hong Kong. In J. Biggs (Ed.), *Testing: To educate or to select? Education in Hong Kong at the crossroads* (pp. 3-12). Hong Kong: Hong Kong Educational Publishing Co.

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

Brown, G. T. L., & Harris, L. R. (2009). Unintended consequences of using tests to improve learning: How improvement-oriented resources engender heightened conceptions of assessment as school accountability. *Journal of Multi-Disciplinary Evaluation, 6*(12), 68-91.

Brown, G. T. L., Kennedy, K. J., Fok, P. K., Chan, J. K. S., & Yu, W. M. (2009). Assessment for improvement: Understanding Hong Kong teachers' conceptions and practices of assessment. *Assessment in Education: Principles, Policy and Practice, 16*(3), 347-363.

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.

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, Singapore.

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

Cheung, T. K.-y. (2008). An assessment blueprint in curriculum reform. *Journal of Quality School Education, 5*, 23-37.

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

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment Research & Evaluation, 10*(7), Available online: http://www.pareonline.net/pdf/v10n17.pdf.

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

Friedman, S. J., & Frisbie, D. A. (1995). The influence of report cards on the validity of grades reported to parents. *Educational and Psychological Measurement, 55*(1), 5-26.

Gable, R. K., & Wolf, M. B. (1993). *Instrument development in the affective domain: Measuring attitudes and values in corporate and school settings.* (2nd ed.). Boston, MA: Kluwer Academic Publishers.

Gao, L. (2010, January). *Teachers' conceptions of assessment: Developing a model for teachers in China.* Paper presented at the annual meeting of the Comparative Education Society of Hong Kong, Guangzhou, China.

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.

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

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

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.

Hui, S. K. F. (2009, August). *Investigating missing conceptions of assessment: Qualitative studies with Hong Kong teachers*. Paper presented at the biannual conference of the European Association for Research in Learning and Instruction.

Hui, S. K. F., Yu, W. M., & Brown, G. T. L. (2010, January). *Teachers' conceptions of assessment: Developing a model for teachers in Hong Kong.* Paper presented at the annual meeting of the Comparative Education Society of Hong Kong, Guangzhou, China.

Kennedy, K. J. (2007, May). *Barriers to innovative school practice: A socio-cultural framework for understanding assessment practices in Asia*. Paper presented at the Redesigning Pedagogy: Culture, Knowledge & Understanding Conference, Singapore.

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.

Li, W. S., & Hui, S. K. F. (2007). Conceptions of assessment of mainland China college lecturers: A technical paper analyzing the Chinese version of CoA-III. *The Asia-Pacific Education Researcher*, *16*(2), 185-198.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

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.

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.

Paine, L. W. (1990). Chinese teachers' views of time. In M. Ben-Peretz & R. Bromme (Eds.), *The nature of time in schools: Theoretical concepts, practitioner perceptions* (pp. 138-157). New York: Teachers College Press.

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

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

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

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

Shohamy, E. (2001). *The Power of Tests: A Critical Perspective on the Uses of Language Tests*. Harlow, UK: Pearson Education.

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.

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

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.

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.

Werner, O., & Campbell, D. T. (1973). Translating, working through interpreters, and the problem of decentering. In R. Naroll & R. Cohen (Eds.), *A Handbook of Method in Cultural Anthropology* (pp. 398-420). New York: Columbia University Press.

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.

Yu, W.M., Kennedy, K.J., Fok, P.K., & Chan, K.S. (2006, May). *Assessment reform in basic education in Hong Kong: The emergence of assessment for learning.* Paper presented at the 32nd Annual Conference of the International Association for Educational Assessment (IAEA), Singapore.

Table 1. Factor EFA Joint Solution

Factor	Item numbers
F1Student Development	14, 29, 20, 24, 22
F2 Irrelevance	23, 31, 26, 10, 49
F3 Examinations	39, 62, 33, 38, 7, 44, 25, 40
F4 Error	36, 58
F5 Help Learning	1, 3, 5
F6 Teacher & School Control	42, 35, 61, 28, 9
F7 Accuracy	60, 12, 6

Table 2. Nested progressively constrained models 2-group confirmatory factor analysis fit statistics

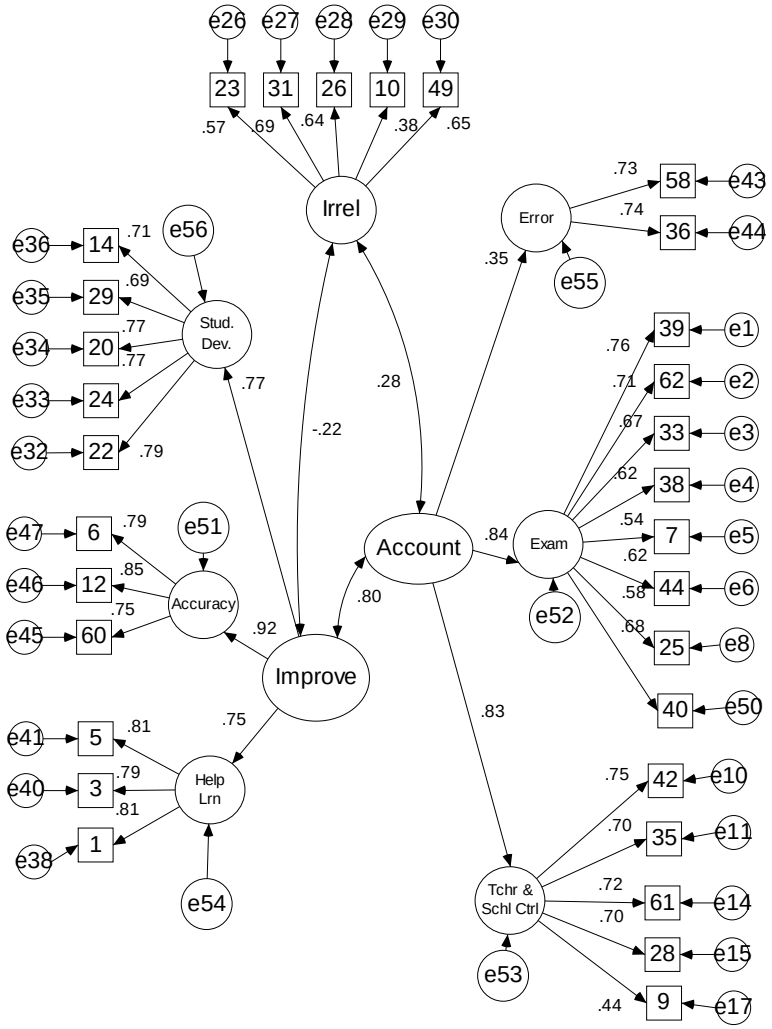
Model	CFI	ΔCFI
Model 0. Unconstrained	.840	—
Model 1. Equivalent item regressions	.839	.001
Model 2. Equivalent factor regressions	.839	.000
Model 3. Equivalent factor covariances	.836	.003
Model 4. Equivalent Factor residuals	.834	.002
Model 5. Equivalent item residuals	.775	.059

Note. CFI=comparative fit index; ΔCFI values <.01 indicate invariance.

Table 3. Descriptive Statistics for C-TCoA Factors in China and Hong Kong

C-TCoA Conceptions & Factors	China				Hong Kong			Effect size (Cohen's *d*)
	k	*M*	*SD*	Cronbach's Alpha	*M*	*SD*	Cronbach's Alpha	
Improvement	*11*	*3.39*	*0.95*	*.89*	*3.68*	*0.73*	*.87*	*-.35*
F1 Student development	5	3.17	1.13	.86	3.30	0.84	.78	-.13
F5 Help learning	3	4.01	1.11	.73	4.65	0.89	.82	-.65
F7 Accuracy	3	3.11	1.13	.78	3.73	0.94	.77	-.60
F2 Irrelevance	*5*	*2.65*	*0.90*	*.64*	*2.28*	*0.76*	*.71*	*.44*
Accountability	*15*	*3.13*	*0.83*	*.84*	*3.40*	*0.83*	*.85*	*-.36*
F3 Examinations	8	3.04	0.97	.83	3.50	0.80	.83	-.53
F4 Error	2	4.23	1.26	.68	4.15	1.01	.64	.07
F6 Teacher & school control	5	2.85	0.99	.74	2.95	0.88	.82	-.11

Note: *k*=number of items; negative values for Cohen's *d* indicate HK is higher, positive values that China is higher. China *n*=898; HK *n*=1014



Note. Values shown are based on Model 4 constrained equivalent model.
Figure 1. Hierarchical 7 Factor Model of Chinese Teacher Conceptions of Assessment

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 Meta Factor

Meta Factor, Factors, and Items

Improvement

F1Student 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。

F5 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。

F7 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

F3 Examinations

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

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

19. Assessment teaches examination-taking techniques.

Meta Factor, Factors, and Items
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。
F4 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。
F6 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.
對學生進行評估所得到的結果可顯示教師是否稱職。
30. Assessment is an accurate indicator of a school's quality.
評估是顯示學校質素的準確指標。
16. Assessment measures the worth or quality of schools.
評估量度學校的價值和質素。
6. Assessment is used by school leaders to police what teachers do.
學校領導利用評估來管治教師所做的事。