The effect of high-stakes examination systems on teacher beliefs: Egyptian Teachers' Conceptions of Assessment

Atta Gebril, *The American University in Cairo, Egypt*

Gavin T L Brown, *The University of Auckland, New Zealand*

Abstract

Egypt is currently attempting to introduce a greater formative use of assessment while maintaining a public examination system. This study investigates teacher beliefs about the purposes of assessment in Egypt, using the Teachers' Conceptions of Assessment (TCoA) inventory. The TCoA inventory elicits responses about four main factors: Improvement, School Accountability, Student Accountability, and Irrelevance. A large sample of (*n*=507) Egyptian pre-service and in-service teachers completed an Arabic version of the TCoA inventory. Confirmatory factor analysis was used to test the Egyptian teachers' responses against the pre-existing New Zealand model, which was found to be inadmissible. Consequently, an ecologically rational 3-factor model was found. The model yielded a strong positive relationship between Improvement and Student Accountability, consistent with previous research. It is argued that greater changes to the examination system are required if teacher beliefs are expected to be more positive about the priority of formative, improvement-oriented uses of assessment.

Accepted for publication in *Assessment in Education: Principles, Policy & Practice*, 2013

Introduction

Tests have always played a gatekeeping role in educational systems where test scores are used for either selection or exit purposes. More recently, assessment has been employed for both school and student accountability within a standards-based framework (Linn, 2000). Especially in examination-driven contexts, this new function has substantially increased the pressures schools face from assessment results. In these contexts, both the quality of a school is judged based on students' scores, as well as the quality of individual students. The No Child Left Behind (NCLB) Act in the US is a very good example of how test scores could be used to judge school quality. US Federal funding is sometimes withheld from schools whose students do not meet the standards stipulated by the NCLB Act. In contrast, public examination systems (e.g., Egypt (Gebril & Hozayen, in press; Gebril & Taha-Thomure, in press) and China (Brown, Hui, Yu, & Kennedy, 2011) use tests to directly judge the quality of students and, since school-based rates of success are published in the public media, indirectly the quality of schools.

Since test- and examination-driven societies depend on teachers to implement or prepare students, it is not possible to discuss assessment issues without considering teachers' roles, responsibilities, and perceptions of assessment practices. Research into teacher beliefs suggests that they act to (1) frame how teachers understand policy innovations, (2) filter out aspects which they consider inappropriate, and (3) guide teachers' responses (Fives & Buehl, 2012). With an increasing emphasis on school-based models of assessment, the role of teachers in student assessment is increasing. Traditionally, teachers are provided with relevant knowledge and skills to meet the demands of new assessment policies. While this training is helpful in principle, it is normally not effective since it does not attempt to understand, or change, implicit beliefs about assessment (Richardson & Placier, 2001). Given the strategic role of beliefs in educational practice, any effort to introduce new assessment practices should consider how teachers conceive of the phenomenon.

The beliefs teachers have about educational activities are expected to be consistent with societal and jurisdictional policy priorities. Teacher beliefs about assessment, teaching, learning, and curriculum have shown strong similarities in societies that prioritise teacher judgement and professionalism as the basis of teacher activity in each of the four domains. Consistent with the notion of ecological rationality (Rieskamp & Reimer, 2007), differences in belief structures and intensities are expected when societies give greater weight to judgements about students, teachers, and schools as a result of publicly administered tests or examinations. In such environments, it is expected that teachers will endorse the societally mandated uses of assessment, even if research studies find such policies and uses not highly associated with enhanced academic achievement.

The current study examines how teachers in Egypt conceive of assessment policy and practices within their high-stakes assessment system. This is achieved by adapting a previously developed self-report inventory and testing its validity with Egyptian teacher education students. Such an approach allows both international comparison of results and the identification of possible differences in teacher beliefs. Together, this design has the potential to identify new areas of research.

Teachers' assessment conceptions

Assessments are used for many purposes (e.g., selection, certification, diagnosis, transfer, licencing, monitoring, etc. described in Newton, 2007). However, it is possible to reduce these multiple uses to four major functions (Brown, 2008). Specifically, assessment can contribute to improved teaching and learning (*Improvement*); assessment can be used to directly evaluate schools and teachers for their effectiveness (*School Accountability*); assessment can certify student achievement, making them accountable for outcomes (*Student Accountability*); and assessment can be considered fundamentally irrelevant to the life and work of teachers and students (*Irrelevant*) (Black, 1998; Heaton, 1975; National Research Council, 2003; Shohamy, 2001; Torrance & Pryor, 1998; Webb, 1992).

While the first three conceptions focus on socially approved functions of assessment, the fourth one focuses on the negative consequences of assessment. The assessment literature often reports a wide range of undesirable effects of tests on teaching and learning, which is usually called negative washback (Alderson & Wall, 1993; Shohamy, 2007). This issue becomes more visible in high-stakes, test-driven educational contexts where 'teaching to the test' practices are widespread, which usually results in narrowing down the curriculum scope (Wall, 2005).

The Teacher Conceptions of Assessment (TCoA) inventory (Brown, 2006) elicits self-reported responses from teachers as to how much they agree with these four main purposes of assessment. The inventory has been widely used as a research tool in multiple countries and languages: English in New Zealand (Brown, 2004, 2011), Queensland (Brown, Lake, & Matters, 2012), and the United States (Calveric, 2010); Chinese in China (Li & Hui, 2007), Hong Kong (Brown et al., 2009); in a variety of European languages—Cyprus (Brown & Michaelides, 2011), Spain (Brown & Remesal, 2012), Netherlands (Segers & Tillema, 2011), and Colombia (Muñoz, Palacio, & Escobar, 2012); and in a range of Islamic societies—Turkey (Vardar, 2010), Iran (Pishghadam, & Shayesteh, 2012), and Pakistan (Khan, 2011). Such widespread interest suggests that the inventory has some efficiency and feasibility as an exploratory method of discerning teacher beliefs about assessment.

Previous studies with the TCoA inventory found statistical invariance between New Zealand primary and secondary teachers (Brown, 2011) and between New Zealand and Queensland primary teachers (Brown, 2006); both jurisdictions give high priority to low-stakes, teacher judgements about performance. A number of studies have recovered instead the four main factors of the TCoA. For example, a Turkish confirmatory factor analysis study recovered the four main factors of the TCoA with middle school teachers (Vardar, 2010). Likewise, in a comparative study of Spanish and New Zealand pre-service teachers the four main factors of the TCoA were recovered (Brown & Remesal, 2012). Similarly, the four main factors were recovered in a study of Greek speaking Cypriot teachers (Brown & Michaelides, 2011).

Nonetheless, in all these 'western' samples, *Improvement* was the most strongly endorsed factor with effect sizes to *Student Accountability* being moderate to large for primary school teachers (d=.50 Queensland, .78 New Zealand) and trivial for high school teachers (d = -.13 Queensland; .12 New Zealand). In contrast, the sample of 103 EFL teachers in Iran endorsed *Student Accountability* more strongly than *Improvement* (d=.34) (Pishghadam & Shayesteh, 2012) and a sample of 414 Turkish middle school teachers endorsed *Student Accountability* the same as *Improvement* (d=.01) (Vardar, 2010). Several studies (Brown et al., 2009; 2011; Li

& Hui, 2007) in Chinese examination systems (i.e., Hong Kong and China) have found high correlations between *Accountability* and *Improvement* purposes, while low inter-correlations were found in jurisdictions with low-stakes, school-based assessments (i.e., New Zealand, Queensland, and Cyprus) (Brown, 2011; Brown, Lake, & Matters, 2011; Brown & Michaelides, 2011).

Teachers' conceptions of assessment in countries with high-stakes examination systems have produced models with quite different characteristics. For example, in China, where teaching and learning are significantly driven by public examinations, Brown, Hui, Yu, and Kennedy (2011) found that accountability is usually perceived not only as indicator of school quality but also as a mechanism for controlling schools, teachers, and students. Also, Chinese teachers perceived both accountability and improvement purposes as highly positively correlated, though accountability was weakly correlated with irrelevance, while improvement was inversely correlated with irrelevance. This result speaks to the conviction of those teachers that examining students is the best way to improve their learning and personal development. This result was considered unusual for Western child-centered pedagogies and was attributed to Chinese cultural values and practices that provide higher status for high-achieving students and which attribute moral virtue to academic success.

Despite differences in assessment policies, these studies suggest that there is some cross-cultural validity for the four main assessment purposes of the TCoA inventory. These differences appear to reflect the contrasting assessment policy priorities and uses of assessment in these quite different societies. Thus, it is expected that teachers working in high-stakes examination societies would have similar responses to the TCoA as exhibited by Chinese teachers.

Assessment context of Egypt

In Egypt, education is dominated by the use of examinations to select students for access to further educational opportunities (Hargreaves, 1997, 2001). Currently, end-of-year exams are the only indicator used in public schools to move student from one educational stage to the next. A very good example of this summative assessment practice is the secondary school leaving exam (*thanaweya amma*). Scores on this test determine which university and academic program students can join. Given the high-stakes nature of the *thanaweya amma* test, students start test preparation almost a year before its administration in June. Test preparation takes a number of forms; the most famous of which is private tutoring. Students start these sessions a month or two before the beginning of the academic year in September. Some families do not send their students to schools during this final year and depend mainly on private tutoring. Statistics have shown that Egyptians spend around 7-8 billion Egyptian pounds on private tutoring every year (around US$1.3 billion). According to a recent publication by the Egyptian Central Agency for Public Mobilization and Statistics (CAPMAS), 42% of household spending on education is allocated to private tutoring. This percentage is even higher than the money spent on tuition school fees, which constitutes 38.8 % of spending on education (CAPMAS, 2013). Newspapers publish expected questions and their exemplar answers to help students before the final exam. The *thanaweya amma* news is usually featured on the front page of most Egyptian newspapers, if not all, during the first three weeks of June. Wide coverage of the reactions of students, parents, and experts about the content of *the thanaweya amma* subject tests is given during this period. The top scorers on the *thanaweya amma* test are invited every

year to meet the minister of education and also offered generous prizes from different officials. Also, schools are awarded when one of their students is on the top list. It is clear from this description that many societal and educational efforts revolve around maximising scores on such public examinations.

The lower grades, at both elementary and preparatory stages, are also dominated by final, high-stakes exams. For example, students with high scores in Grade 6 are placed in better schools once they move to from the primary stage to preparatory schools (starting from grade 7). Also, based on their scores on the Grade 9 final exam, students are placed either in the more prestigious general secondary schools or sent to technical / vocational schools (Lloyd, El Tawila, Clark, & Mensch, 2003). Generally, high achievers in Egypt are well-respected in their families and also in their schools. Hargreaves (1997, p. 167) provides an excerpt from an interview with a school kid that shows the overriding importance of performance on tests. In this interview, a student describes one of her colleagues who failed one of the tests: "Her friends hated her because she failed twice . . . I don't hate her, but I don't want to be like her. She failed because she didn't study all the time." This quotation shows how human relationships in schools are shaped by examinations. In the same article, Hargreaves also argues that "examination orientation in the Egyptian school was all about winning the prize at the end rather than the intrinsic satisfaction of running the race" (p. 168). This ideology has been deeply rooted even in the Egyptian bureaucracy where students graduating with the highest GPA in each academic program are automatically employed by the government since they receive a permanent full-time job.

In an attempt to provide a balance between summative large-scale testing and formative assessment, the Ministry of Education introduced the Comprehensive Assessment (CA) initiative. CA attempts to embed assessment activities within instruction and make it an ongoing process. The following quotation from the 2007-2012 strategic plan describes this process:

> Recently, attempts have been made to introduce change at the early primary level in the form of a combined assessment approach, based on National Standards, namely Comprehensive Assessment. The final grade in primary grades one through three is based on an exam score combined with performance on activities and an on-going student portfolio. It is hoped that this model will serve as a prototype for reform of assessment method at higher levels (Egyptian Ministry of Education, 2007, p. 44).

As described in the previous paragraph, it is clear that the CA initiative attempts to move away from sole dependence on traditional testing in primary classes. Also, CA is perceived as a first step in a reform process that would be extended to later educational stages. It is envisioned that this reform process focuses on:

- making assessment an integral part of learning,
- making assessment an ongoing process,
- using alternative assessment tools along with traditional exams,
- promoting critical thinking and problem-solving strategies essential for life-long learning, and
- developing national assessment standards.

The CA policy was initially implemented at lower grades in elementary schools (Grades 1- 3) in 2005. By 2010, the policy was adopted from in all elementary and

preparatory public schools (Grades 1-9). Future plans include implementing the policy in at the secondary stage.

This improvement-oriented policy seems to differ from the dominant examination practices that emphasize the accountability and selection functions of tests. For this reason, it is very important for policy makers to provide resources so that teachers have a better understanding of the policy and capacity to implement it. After the January 25th Revolution, there is cautious optimism amongst different stakeholders in Egypt. The new government is looking into teachers' salaries and work conditions. However, since the economy has not recovered yet, the implementation of the new policies will not be effective soon. On the academic side, teachers need to be equipped with the knowledge and skills pertaining to how and why they should use CA. In addition, we need to ensure that teachers have positive beliefs about this policy. Hence, it is important to investigate how assessment is perceived by teachers, the focus of this study. Generally, it is expected that Egyptian teachers will have beliefs that are consistent with the Egyptian uses of assessment; that is, they will believe that the main purpose of assessment is making students accountable and secondly, they will endorse the notion that assessment serves the purpose of improved teaching and learning.

Research Questions

The goal of this study then was to examine the beliefs of Egyptian teachers about the purposes of assessment, making use of the Teachers' Conceptions of Assessment (TCoA) inventory. Since the CA system is still relatively new to Egypt, it was decided to investigate the beliefs of pre- and in-service teachers currently enrolled in education programs. Focusing on this sample meant that it was more likely positive views about formative assessment would be elicited. The following research questions were set:

1. Would an Arabic translated version of the TCoA elicit responses that fit the New Zealand statistical model (i.e., four main factors in a hierarchical, inter-correlated structure) and, if not, which model would best represent the beliefs of Egyptian teachers?
2. What do the factors mean scores and inter-correlations suggest concerning the conceptions of assessment held by Egyptian teachers?
3. What modifications to the TCoA, if any, would be warranted for future research with practicing teachers in Egypt?

Methodology

Research Instrument

The Teacher Conceptions of Assessment version III abridged (TCoA) inventory, consisting of 27 items (Brown, 2006) was used. Responses to the TCoA aggregate into nine factors, each made up of three items, which form four inter-correlated, intention-oriented conceptions of assessment (i.e., *Improvement*, *Student Accountability*, *School Accountability*, and *Irrelevance*). The *Improvement* conception has four contributing factors (i.e., assessment **describes** student learning, assessment is **valid**, assessment improves **student learning**, and assessment improves **teaching**). In other words, *Improvement* depends upon assessment providing valid and accurate descriptions of learning, as well as guiding students and teachers on how to improve. Similarly, *Irrelevance* has three contributing factors (i.e., assessment is **unfair**, assessment is **ignored**, and

assessment is **inaccurate**). This meant that the *Irrelevance* conception indicates that assessment is bad for students, is ignored by teachers, and is inaccurate.

Translation and Adaptation

The TCoA abridged questionnaire (Brown, 2006) was translated into Arabic by one of the researchers. The questionnaire includes 27 items and also a number of demographic questions. We adopted a functional equivalence rather than a literal translation approach when we translated the items into Arabic. In order to check the quality of the translated version, we followed a number of procedures. First, two research assistants, native speakers of Arabic, read the translated questionnaire closely, and discussed possible changes with one of the researchers. Then, the questionnaire was submitted to both language and assessment specialists who also gave feedback on the quality of the different items in terms of both linguistic and content-related issues. The final draft of the questionnaire was piloted with a group of teachers to check its clarity and appropriateness. Based on feedback from research assistants, experts, and teachers, the questionnaire was revised and some modifications were made. For example, some of the statistical terms were rephrased in order to prevent any confusion on the part of teachers since most teachers are not familiar with these technical concepts. We used familiar vocabulary and sometimes explanatory statements to clarify assessment-related terminology like measurement error and imprecision. Also, we systematically used the word "تقييم" (a term commonly used to refer to assessment) as opposed to "تقويم" (a term employed to refer to evaluation). Before completing the questionnaire, the participating teachers were informed about these issues in order to avoid any misunderstanding.

The TCoA uses a positively packed agreement rating scale in which there are two negative and four positive options. Such scales have been found to be effective (Brown, 2004; Lam & Klockars, 1982) when participants are inclined to agree with all constructs. Tendency to agree may be more prevalent when teachers are expected to implement and thus, agree with, government policies.

Participants

Participants in this study were selected from pre-service and in-service teachers enrolled in a public university in one of the Southern regions of Egypt (Table 1). The pre-service teachers participating in this project were undergraduate students who were in their final year of classes. Those students were preparing to work in Egyptian public schools and had teaching practice experience in public schools near the university. As for the in-service sample, those were teachers with full-time jobs in public schools while pursuing a diploma in education by attending evening classes in the same university where data were collected. The educational diploma is one of the requirements for promotion in public schools and consequently an increasing number of teachers are enrolling in this program in colleges of education country-wide. Most of these participants had less than five year experience, though one individual had over 25 years of service. Unsurprisingly, the vast majority of participants were female. There was greater diversity of teaching specialisation among the pre-service participants than among the in-service teachers.

Participants were asked to self-rate their assessment literacy and teaching competence on a 10-point scale; results were very similar and reasonably optimistic. However, assessment literacy self-ratings were lower than teaching competency evaluations by sizeable margins (Cohen's d=.57 pre-service, .53 in-service).

Table 1. Demographic characteristics of participants

Characteristic	Pre-service teachers (*n*=305)	Inservice teachers (*n*=202)
Sex		
Female	271	162
Male	34	40
Subject Specialization		
English	70	100
Math & science	72	30
Arabic	80	34
Kindergarten	83	0
Other	0	38
Teaching experience	—	74% <5 years
	M (*SD*)	*M* (*SD*
Self-reported assessment literacy	6.41 (1.47)	6.53 (1.54)
Self-reported teaching competency	7.24 (1.43)	7.32 (1.42)

Data collection

Before data collection, we contacted a number of professors in the target the university to look into the possibility of recruiting students for this project. Those professors kindly agreed to ask their students to participate in this study. Based on this initial approval, we contacted the college of education administration and obtained final permission to collect data from intact classes. This strategy helped the researchers collect data from a relatively large number of participants. One of the researchers visited these classes and gave some background information about the project. Also, the participants were informed about the purpose of the study and were given a consent form to sign. In total, the participants completed over 550 questionnaires, but the researchers decided to remove around 50 incomplete questionnaires from the data. Five hundred and seven questionnaires were included in the final data set.

Data Analysis

Confirmatory factor analysis, in Amos v.20, was used to test Egyptian teacher responses against the New Zealand model. Rather than develop a completely new model with exploratory factor analysis, modifications to the original solution were attempted to maximise comparability of results to previous studies. Even with large sample sizes (Boomsma & Hoogland, 2001), improper solutions (e.g., non-positive definite covariance matrices or negative error variances) can occur. Since small sample size is not a valid explanation for such results, it was considered that structural modifications were needed to identify the structure of this sample of participants. Valid solutions for improper results include removing 1st-order factors (e.g., within the *Irrelevance* and *Improvement* TCoA factors) so that items load directly onto 2nd-order factors, fixing negative error variances to a small positive value if two times the standard error is larger than the observed value, or converting a correlation (e.g., among any two of the four TCoA factors) into a linear dependency so that fewer inter-correlated factors existed while maintaining factors (Chen, Bollen, Paxton, Curran, & Kirby, 2001).

After modifications, the resulting Egyptian TCoA model was tested for fit using confirmatory factor analysis; technically, this is a restrictive analysis because

no new data were collected. Non-rejection of the confirmatory model was determined if the χ^2/df ratio had $p>.05$ (Marsh, Hau, Wen, 2004), gamma hat >.90 and SRMR ≈ or <.06 (Fan & Sivo, 2007), and RMSEA <.08 (Hu & Bentler, 1999).

In order to test the invariance of the trimmed model for the pre- and in-service teacher groups, nested multi-group confirmatory factor analysis was used. This approach tests the difference in the fit of a model as a consequence of sequentially constraining parameters in the model to be equivalent between groups. Differences in the comparative fit index of <.01 indicate that the constraining process has resulted in a statistically equivalent parameter (Cheung & Rensvold, 2002). The parameters tested for the TCoA model are equivalent regression weights from factors to items, equivalent regression weights from 2nd-order factors to 1st-order factors, equivalent correlations between factors, equivalent factor residuals, and equivalent item residuals. If the model is equivalent (except for item residuals), then it is taken that the groups are drawn from the same population and that any differences in factor means are a function of real differences, rather than a consequence of the inventory (Wu, Li, & Zumbo, 2007).

Results

The New Zealand model was inadmissible because of negative error variance on the 1st-order *Describe* factor and a correlation >1.00 between *Improvement* and *Student Accountability*. To resolve these, the model was modified so that the items for *Describe* were loaded directly onto *Improvement* and so that the *Student Accountability* factor became a subordinate 1st-order factor to *Improvement*, resulting in inter-correlations only among *Improvement*, *Irrelevance*, and *School Accountability*. This model had a negative error variance smaller than two times the standard error on *Student Accountability* and so was fixed to .005. All other items and factors were retained as per the TCoA model (Figure 1); the model met criteria for good fit to the data (N=507; χ^2=833.58; df=314; χ^2/df=2.655, p=.10; CFI=.88; gamma hat =.93; RMSEA=.057, 90%CI=.052-.062; SRMR=.0545). Nested invariance testing for group showed that the difference in CFI was <.01 for each parameter constraint up to and including item residuals; fit of this fully constrained model also met expectations (χ^2=1303.10; df=695; χ^2/df=1.875, p=.17; CFI=.86; gamma hat =.96; RMSEA=.042, 90%CI=.038-.045; SRMR=.0711).

As can be seen in Figure 1, *Improvement* strongly predicted the original four 1st-order factors within the construct and the *Student Accountability* factor. This indicates clearly that improving student learning and teacher's instruction is bundled with evaluating students. Likewise, the *Irrelevance* factor predicted strongly the original three 1st-order factors, and the *School Accountability* factor strongly predicted the original three items for this factor.

While the two-group model had good fit to the data, five of the factor to item loadings were <.30 (i.e., one in **ignore**, two in **inaccuracy**, one in *Student Accountability*, and one in **valid**). These items, despite translation quality, clearly do not reflect the intended constructs among this sample of Egyptian teachers. Looking at Appendix A, three of these items (items 9, 15, and 18) are associated with score reliability issues. Those teachers may not be familiar with these technical aspects of reliability and, consequently, there is a great possibility that they misunderstood the content of these statements.

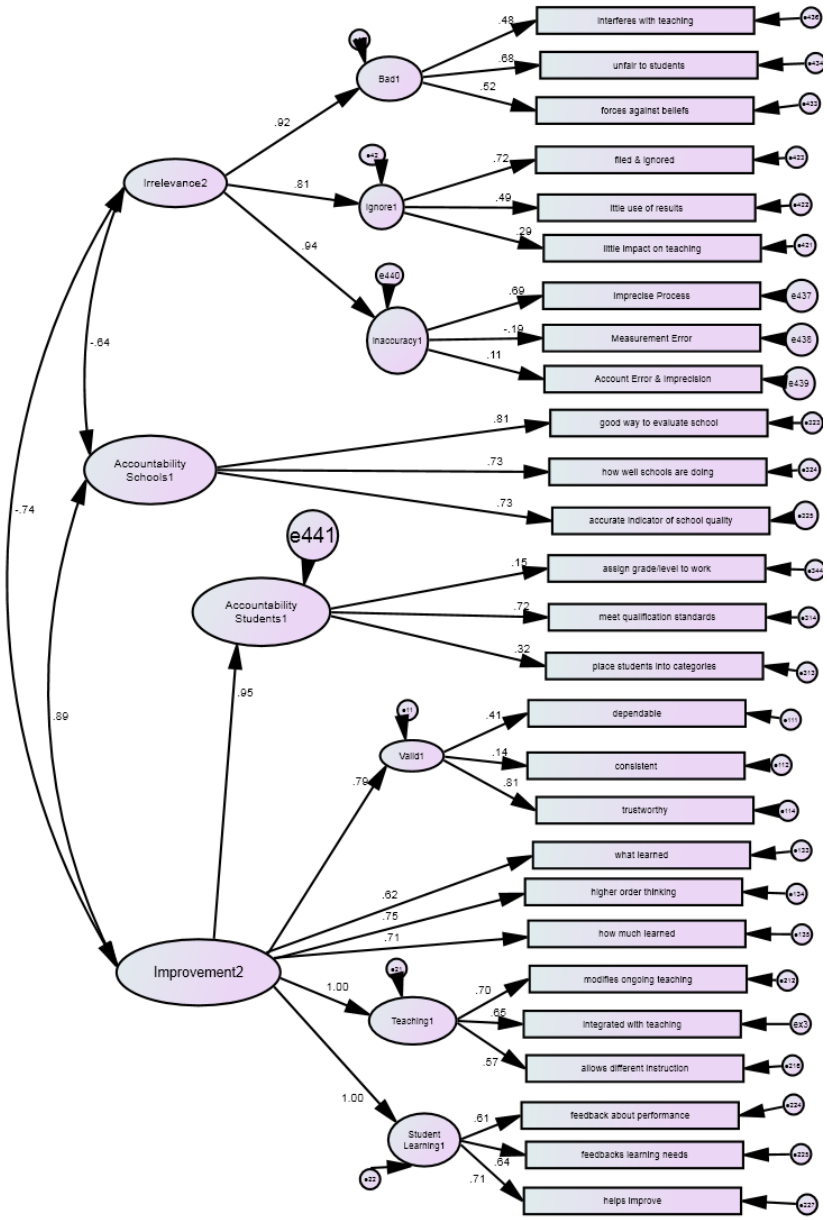

Figure 1. Egyptian TCoA statistical model

Based on this model, scale scores were created for the three main factors and mean score differences were examined for the two groups of teachers (Table 2). The mean scores all fell between slightly and moderately agree and the only statistically significant difference was for *Improvement* which in-service teachers endorsed more than pre-service teachers; however, the difference is still only moderate at best.

Table 2. Egyptian TCoA factor inter-correlations and scale values by teacher type

	Inter-correlations			Pre-Service		In-Service		Differences	
TCoA Factors	1.	2.	3.	*M*	*SD*	*M*	*SD*	*F*	Cohen's *d*
1. Improvement	—			3.66	.86	3.92	.81	11.56	-.31
2. School Accountability	.89	—		3.51	1.22	3.45	1.28	.22	.05
3. Irrelevance	-.74	-.64	—	3.65	.79	3.65	.80	.002	.00

Discussion

This study has shown that the original factors of the TCoA could be recovered in this sample of teacher education students in Egypt. However, the structural paths among the factors were both statistically and practically divergent from the original model developed in New Zealand. Hence, there is some validation support for the TCoA factors and items but not the structural relations found elsewhere. The current results suggest that considerable work is needed to understand Egyptian teacher responses to the negative aspects of assessment; perhaps, Egyptian assessments are not ignored and are not considered inaccurate or invalid. Further research in Egypt is needed to understand why these items do not behave as expected and why the factors have different relationships. Such research should identify whether there are other beliefs about assessment not even addressed by the TCoA, as well as developing better indicators for the under-represented constructs.

The striking characteristic of the revised model is the strong association between *Improvement* and *School Accountability* factors. Within an examination-driven society, it seems entirely logical and rational to link examination scores with the quality of schools and educational improvement. It is as if teachers believe "*a high quality school improves student learning and teaching which shows in high scores on examinations; judge us by our student exam results"*. Since there is evidence that the current Egyptian assessment regime operates this way, we should not be surprised that teachers' beliefs conform. As indicated earlier, Egyptian schools take pride in having a high percentage of students passing the *thanawiya amma* test. Also, schools are rewarded when one of their students is among highest-scoring test takers on this exam nation-wide.

This strong agreement aligns with previous research in high-stakes examination contexts (Brown et al., 2011; Brown, Kennedy, Fok, Chan, & Yu, 2009). As argued by Brown et al. (2011), this result is expected in such an examination-driven environment where teachers believe that tests lead to better learning and enhance student motivation. Needless to say, the Egyptian teachers do not share Confucian philosophy or Chinese imperial examination history that are often used to explain the strong pro-examination beliefs of Chinese teachers. It is possible that Egyptian assessment beliefs are a consequence of British colonisation in the 20th century. Mansfield (1971, p. 139) argues that "no aspect of the British occupation of Egypt is more open to criticism than its effect on education". More speculatively, it is possible that the traditional Islamic education that has always promoted memorization of religious texts has had some influence on Egyptian approaches to assessment. Another reason could be the dramatic increase of school enrolment during Nasser's time (1952-1970) since free education was granted to all

students. As argued by Radwan (cited in Douara, 2008, p. 14), the " unfortunate consequence of expansion in education has been the directing of the educational progress towards academic achievement more than towards training students for the ability to undertake research ad practice thinking." A final interpretation could be the centralized system of education in Egypt which has always used exams as a way to control schools, students, and teachers; and consequently has created an environment where evaluating students is perceived as a proper part of improvement. Thus, this study provides further evidence that beliefs in high-stakes examination societies have strong similarities. Nonetheless, future research about assessment beliefs and practices in Arabic and Islamic societies would benefit from a more clearly delineated understanding of societal factors that privilege the high-stakes, examination systems so widely practiced in those societies. More qualitative, intensive research with Egyptian teachers, teacher education students, school leaders, and policy makers is clearly warranted to better understand the results reported in this paper.

It is actually tempting to use a test-driven strategy (Popham, 1993; Torrance, 2009, 2011) for improving educational practices in Egypt given its centralized system of education, which, in turn, allows for quick nation-wide implementation of any new policy. However, there are a number of pre-conditions that should exist to ensure its success and sustainability. The current policy push for an assessment-for-learning model (i.e., Comprehensive assessment initiative), while having some merit, is clearly failing to modify the dominant belief system of teachers who construe improvement as a means of holding students accountable. Using assessments in a strongly formative way requires reducing the emphasis on assessment as an evaluative tool in favour of a more diagnostic approach. Further, the focus of formative assessment has to move from presuming that it is the student who must change in order to improve, to an approach that encourages teachers to modify their instruction so as to cause learning to take place. This shift is difficult to implement as long as examinations are maintained as the 'real' system of evaluating success. Thus, it seems highly likely that the current CA policy is failing, in part, because the standards based assessments are a soft option relative to the hard policy option that continues the use of examinations even at the end of Grade 3 (Kennedy, Chan, & Fok, 2011). A much braver reform effort would remove all examinations within primary schooling so that all students have free access to high school education opportunities. As long as a traditional mind-set exists (i.e., poor performance is a function of the learner's poor motivation and effort) and policies privilege high-stakes examinations, it is unlikely that any formative assessment will actually be used to guide improved or changed teaching.

On the presumption that removal of examinations is unlikely, this study suggests that any attempt to reform Egyptian assessment practices is likely to fail, unless efforts focus on using formal tests as a tool for diagnostic analysis of student learning. Carless (2011) has shown that Hong Kong teachers can be taught to use their regular summative testing in a diagnostic formative fashion to change their instructional practices in response to identified performance problems. Furthermore, given the difficulty in helping teachers to create diagnostic tests, Egypt may benefit from adopting the example of New Zealand where IT-supported formative testing systems (Archer & Brown, 2013; Brown & Hattie, 2012; Hattie & Brown, 2008) have led to improved teaching practices and learning outcomes (Lai, McNaughton, Timperley, & Hsiao, 2009; Parr & Timperley, 2008).

The in-service teachers were somewhat more committed to the notion of improvement as the purpose of assessment. The logic behind this endorsement may simply reflect compliance with an official policy or else recognition of the *force majeure* of assessment—teachers are judged by society for student results. However, this greater commitment is a potential lever for helping teachers use new methods of formative assessment. Nonetheless, this too is an area for further research; disentangling real commitment from social desirability responses is an important task in teacher development.

While it is important that teachers develop beliefs about the purposes of assessment that lead to improved outcomes, the Ministry of Education should not presume that teacher beliefs are the problem. If beliefs are ecologically rational, then the Ministry probably needs to change the conditions and consequences of assessment to expect a change in attitudes, values, or beliefs. As argued in New Zealand (Brown & Hattie, 2012; Hattie & Brown, 2008, 2010), consequences for poor performance have to be reduced if there is to be any chance that teachers will use assessment results to consider the possibility that their own teaching practices might need to change in order to permit greater or deeper learning to take place. Assessment literacy programs for teachers that aim at changing teacher's perception of assessment and providing them with appropriate technical knowledge and skills are certainly required. But unless all social uses and effects of assessment are considered, it is unlikely that teachers' beliefs will change. Hence, the current results clearly identify where government policy needs to focus its attention: formative assessment with low consequences. A key to success of such an approach is the sustained appointment of personnel with appropriate values and skills (Kuan, 2011), a problem in highly politicised institutional environments.

Conclusions

This study has found that the original statistical model for the Teacher Conceptions of Assessment inventory did not fit Egyptian teachers' responses. A revised model that was ecologically rational with Egypt's high-stakes public examination system showed that three distinct purposes for assessment could be identified (i.e., improvement, school accountability, and irrelevance) and that improvement was highly associated with evaluating students. While new items are needed to more fully understand teacher conceptions of assessment in Egypt, the current study has shown that Egyptian teachers' conceptions are more alike to those of Chinese and Hong Kong teachers who also work in high-stakes, public examination systems than the beliefs of teachers in New Zealand and Queensland who work in low-stakes, formative assessment systems. This study adds to our understanding of the impact of accountability and contexts on teacher beliefs.

Acknowledgments:

The researchers would like to thank the research assistants who helped in translating the questionnaire. Also, our appreciation is extended to the experts who gave feedback on an early draft of the translated questionnaire and to our colleague Dr. Mohamed Ismail for his efforts in recruiting participants and also for helping in data collection. This study could have been possible without the help of the teachers who provided the study data.

Notes about contributors:

Accepted in *Assessment in Education: Principles, Policy & Practice*, 2013

Atta Gebril is an assistant professor in the MATESOL program at the American University in Cairo, where he teaches courses in language assessment and research methods. He obtained his PhD in foreign language and ESL education with a minor in language testing from the University of Iowa. His research interests include writing assessment, reading–writing connections, test validation, and teacher beliefs. His work has appeared in a number of journals including *Language Testing, Assessing Writing, and Language Assessment Quarterly*. He has taught in the US, United Arab Emirates, and Egypt. Also, he has worked on a number of test development projects in different parts of the world. Email: agebril@aucegypt.edu

Gavin T. L. Brown is Associate Professor of Education at The University of Auckland, New Zealand. Gavin has published over 80 research articles in refereed journals and book chapters and written two textbooks on assessment, including*Contemporary educational assessment: Practices, principles, and policies* (Pearson South Asia, 2010) and *An introduction to educational assessment, measurement, and evaluation: Improving the quality of teacher-based assessment* (Pearson Education New Zealand 2nd edition, 2008). He is also a co-author of two standardised educational test systems published in New Zealand, including *Essential Skills Assessment: Information Skills* (NZCER, 2001) and *Assessment Tools for Teaching and Learning (asTTle)* (Ministry of Education, 1st to 4th editions, 2002-2005). Gavin's major research interest is cross-cultural study of the social psychological effects of assessment on prospective and in-service teachers and upon school and higher education students. E-mail: gt.brown@auckland.ac.nz

References

Alderson, C., & Wall, D. (1993). Does washback exist? *Applied linguistics, 14*(2). 115–129.

Archer, E., & Brown, G. T. L. (2013). Beyond rhetoric: Leveraging learning from New Zealand's Assessment Tools for Teaching and Learning for South Africa. *Education as Change: Journal of Curriculum Research, 17*(1), 131-147. doi:10.1080/16823206.2013.773932

Boomsma, A., & Hoogland, J. J. (2001). The robustness of LISREL modeling revisited. In R. Cudeck, S. Du Toit & D. Sorbom (Eds.), *Structural equation modeling: Present and future* (pp. 139-168). Lincolnwood, IL: Scientific Software International.

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

Brown, G. T. L. (2011). Teachers' conceptions of assessment: Comparing primary and secondary teachers in New Zealand. *Assessment Matters, 3*, 45-70.

Brown, G. T. L., & Hattie, J. A. (2012). The benefits of regular standardized assessment in childhood education: Guiding improved instruction and learning. In S. Suggate & E. Reese (Eds.), *Contemporary debates in child development and education* (pp. 287-292). London: Routledge.

Brown, G. T. L., & Michaelides, M. P. (2011). Ecological rationality in teachers' conceptions of assessment across samples from Cyprus and New Zealand. *European Journal of Psychology of Education, 26*(3), 319-337. doi:10.1007/s10212-010-0052-3

Accepted in *Assessment in Education: Principles, Policy & Practice*, 2013

Brown, G. T. L., & Remesal, A. (2012). Prospective teachers' conceptions of assessment: A cross-cultural comparison. *The Spanish Journal of Psychology, 15*(1), 75-89. doi:10.5209/rev_SJOP.2012.v15.n1.37286

Brown, G. T. L., Kennedy, K. J., Fok, P. K., Chan, J. K. S., & Yu, W. M. (2009). Assessment for student improvement: understanding Hong Kong teachers' conceptions and practices of assessment. *Assessment in Education: Principles, Policy & Practice, 16*(3), 347-363. doi:10.1080/09695940903319737

Brown, G. T. L., Lake, R., & Matters, G. (2011). Queensland teachers' conceptions of assessment: The impact of policy priorities on teacher attitudes. *Teaching and Teacher Education, 27*(1), 210-220. doi:10.1016/j.tate.2010.08.003

Calveric, S. B. (2010). *Elementary Teachers' Assessment Beliefs and Practices* (doctoral dissertation). Virginia Commonwealth University, Richmond, VA.

Carless, D. (2011). *From testing to productive student learning: Implementing formative assessment in Confucian-Heritage settings*. London: Routledge.

Central Agency for Public Mobilization and Statistics(CAPMAS) – Egypt. (2013). *Indicators of income, expenditure, and consumptions*. Retrieved from http://capmas.gov.eg/pdf/studies/pdf/enf2012.pdf

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

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

Douara, D. M. S. (2008). *Rote learning in the Egyptian national education system: Possible roots and consequence* (MA thesis). American University in Cairo, Egypt.

Egyptian Ministry of Education. *2007-2102 stretegic plan*. Cairo: Ministry of Education.

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

Fives, H., & Buehl, M. M. (2012). Spring cleaning for the "messy" construct of teachers' beliefs: What are they? Which have been examined? What can they tell us? In K. R. Harris, S. Graham & T. Urdan (Eds.), *APA Educational Psychology Handbook: Individual Differences and Cultural and Contextual Factors* (Vol. 2, pp. 471-499). Washington, DC: APA.

Gebril, A. & Hozayin, R. (in press). Assessing English in the Middle East and North Africa. In Antony Kunnan (Ed.), *The companion to Language Assessment*. Malden, MA: Wiley-Blackwell.

Gebril, A. & Taha-Tamure, H. (in press). L1/L2 Arabic assessment. In Antony Kunnan (Ed.), *The companion to Language Assessment (*pp. xxx-xxx). Malden, MA: Wiley-Blackwell

Hargreaves, E. (1997). The diploma disease in Egypt: Learning, teaching and the monster of the secondary leaving certificate. *Assessment in Education: Principles, Policy & Practice, 4*(1), 161-176. doi:10.1080/0969594970040111

Hargreaves, E. (2001). Assessment in Egypt. *Assessment in Education: Principles, Policy & Practice, 8*(2), 247-260. doi**:**10.1080/09695940124261

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

Hattie, J. A., & Brown, G. T. L. (2010). Assessment and evaluation. In C. Rubie-Davies (Ed.) *Educational psychology: Concepts, research and challenges* (pp. 102-117). Abingdon, UK: Routledge.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55.

Khan, A. (2011). *Secondary school mathematics teachers conceptions regarding assessment* (Unpublished master's dissertation). Aga Khan University, Karachi, Pakistan. Retrieved from http://ecommons.aku.edu/theses_dissertations/415

Kennedy, K. J., Chan, J. K. S., & Fok, P. K. (2011). Holding policy-makers to account: exploring 'soft' and 'hard' policy and the implications for curriculum reform. *London Review of Education, 9*(1), 41-54. doi:10.1080/14748460.2011.550433

Kuan, L. (2011) . *EQUIP2 lessons learned in education: Student assessment*. Washington DC: USAID

Lai, M. K., McNaughton, S., Timperley, H., & Hsiao, S. (2009). Sustaining continued acceleration in reading comprehension achievement following an intervention. *Educational Assessment, Evaluation and Accountability, 21*(1), 81-100.

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.

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

Lloyd, C, El Tawila, S., Clark, W., & Mensch, B. (2003) The impact of educational quality on school exit in Egypt. *Comparative Education Review*, 47(4), 444-467.

Mansfield, P. (1971). *The British in Egypt*. London: Cox & Wyman Ltd.

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.

Muñoz, A. P., Palacio, M., & Escobar, L. (2012). Teachers' Beliefs About Assessment in an EFL Context in Colombia. *PROFILE, 14*(1), 143-158.

Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy & Practice, 14*(2), 149-170.

Parr, J. M., & Timperley, H. (2008). Teachers, schools and using evidence: Considerations of preparedness. *Assessment in Education: Principles, Policy & Practice, 15*(1), 57-71.

Pishghadam, R., & Shayesteh, S. (2012). Conceptions of Assessment among Iranian EFL Teachers. *The Iranian EFL Journal, 8*(3), 9-23.

Popham, W. J. (1993). Measurement-driven instruction as a" quick-fix" reform strategy. *Measurement and Evaluation in Counseling and Development, 26*(1), 31-34.

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

Rieskamp, J., & Reimer, T. (2007). Ecological rationality. In R. F. Baumeister & K. D. Vohs (Eds.), *Encyclopedia of Social Psychology* (pp. 273-275). Thousand Oaks, CA: Sage.

Segers, M., & Tillema, H. (2011). How do Dutch secondary teachers and students conceive the purpose of assessment? *Studies in Educational Evaluation, 37*(1), 49-54. doi:10.1016/j.stueduc.2011.03.008

Shohamy, E. (2007). The power of language tests, the power of the English language and the role of ELT. In J. Cummins & C. Davison (Eds.), *International Handbook of English Language Teaching* (pp. 521–531). New York: Springer.

Torrance, H. (2011). Using assessment to drive the reform of schooling: Time to stop pursuing the chimera? *British Journal of Educational Studies, 59*(4), 459-485.

Vardar, E. (2010). *Sixth, seventh and eighth grade teachers' conception of assessment* (Unpublished masters thesis). Middle East Technical University, Ankara, Turkey.

Wall, D. (2005). *The impact of high-stakes examinations on classroom teaching: A case study using insights from testing and innovation.* Cambridge: Cambridge 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.

Appendix A: Teachers' conceptions of assessment questionnaire

1. Assessment provides information on how well schools are doing
يوفر التقييم معلومات عن مدى إجادة المدارس فى أداء مهامها

2. Assessment places students into categories
يصنف التقييم الطلاب إلى فئات

3. Assessment is a way to determine how much students have learned from teaching
يعد التقييم طريقه لتحديد مدى استفادة الطلاب من التدريس

4. Assessment provides feedback to students about their performance
يزود التقييم الطلاب بمعلومات عن أداءهم

5. Assessment is integrated with teaching practice
هناك تكامل بين التقييم والممارسات التدريسية

6. Assessment results are trustworthy
نتائج التقييم جديرة بالثقة

7. Assessment forces teachers to teach in a way against their beliefs
يفرض التقييم على المعلمين التدريس بطريقه تتعارض مع قناعاتهم

8. Teachers conduct assessments but make little use of the results
يقوم المعلمون بإجراء التقييم ولكنهم قليلا ما يستفيدون من نتائجه

9. Assessment results should be treated cautiously because of measurement error
يجب أن تعامل نتائج التقييم بعناية شديدة بسبب أخطاء القياس الموجودة فى هذه النتائج

10. Assessment is an accurate indicator of a school's quality
التقييم هو مؤشر دقيق لجودة المدرسة

11. Assessment is assigning a grade or level to student work
التقييم هو وضع درجة أو مستوى لعمل الطالب

12. Assessment establishes what students have learned
يرسخ التقييم ما قام الطلاب بدراسته

13. Assessment feeds back to students their learning needs
يؤدى استخدام التقييم إلي إدراك الطلاب لحاجاتهم التعليمية

14. Assessment information modifies ongoing teaching of students
تساعد المعلومات المستقاة من التقييم على تعديل طرق التدريس المستخدمة حالياً

15. Assessment results are consistent

نتائج التقييم متسقة دوما

16. Assessment is unfair to students

التقييم غير منصف للطلاب

17. Assessment results are filed & ignored

توضع نتائج التقييم فى ملفات و تهمل

18. Teachers should take into account the error and imprecision in all assessment

يجب على المعلمين الأخذ في الاعتبار عدم دقة نتائج التقييم (و احتمال وجود نسبة خطأ)

19. Assessment is a good way to evaluate a school

يعد التقييم طريقه جيده لقياس مدى كفاءة المدرسة

20. Assessment determines if students meet qualifications standards

يحدد التقييم مدى توافق الطلاب مع معايير الكفاءة

21. Assessment measures students' higher order thinking skills

يقيس التقييم مهارات التفكير العليا لدى الطلاب

22. Assessment helps students improve their learning

يساعد تقييم الطلاب على تحسين مستواهم التعليمى

23. Assessment allows different students to get different instruction

يسمح التقييم لمختلف الطلاب في الحصول على تعليم مناسب لكل منهم على حده

24. Assessment results can be depended on

يمكن الاعتماد على نتائج التقييم

25. Assessment interferes with teaching

يتعارض التقييم مع التدريس

26. Assessment has little impact on teaching

للتقييم تأثير ضئيل على التدريس

27. Assessment is an imprecise process

يعد التقييم عمليه غير دقيقة