Conceptions of Assessment: Investigating What Assessment Means to Secondary and University Teachers

Arab Journal of Applied Linguistics
e-ISSN 2490-4198
Vol. 1, No. 1, January 2016, 19-43
© The Author
http://www.arjals.com

Sahbi Hidri
Faculty of Human and Social Sciences of Tunis, Tunisia

Abstract

The overriding objective of this study was to investigate secondary and university teachers' assessment conceptions in an EFL context using a four-factor (*Student Accountability, School Accountability, Improvement* and *Irrelevance*)teachers' conceptions of assessment (TCoA) inventory (Brown, 2006). Data of the study were collected by a questionnaire administered to secondary school (*n*=336) and university (*n*=206) teachers.Factor analyses (exploratory factor analysis (EFA), parallel analysis (principal component analysis (PCA)), dimension analysis (SPSS R-Menu v.2.0) and confirmatory factor analysis (CFA)) using SPSS v. 22.0 and AMOS v. 21.0 were examined. Results delineated a three-factor model (*Accountability*, *Improvement* and *Irrelevant*)withan endorsement of and a significant relationship between *Accountability* and *Improvement*. The ecological relationship between factors and indicators denoted teachers' misconceptions about assessment. Implications for future research on TCoA in similar-related contextswere also discussed.

Keywords: *Assessment literacy, TCoA, education policy,EFA, PCA, CFA, assessment misconceptions*

Background to the Study

One of the major roles ofassessment conceptions is to preserve ethics and standards (Conley, 2005; Davies, 2008a). Such conceptions impact teachers' views on how to conceive of learning, teaching and assessment intended to cope with instructional objectives and learning outcomes; a trend that was ignoredduring the early days when researchers (e.g., Lado, 1961) started theorizing about language testing. Many researchers (e.g., Inbar-Lourie, 2008; Malone, 2013; Stiggins, 2002; Stoynoff & Chapelle, 2005) have accentuated varying definitions of assessment literacy. According to Taylor (2009), assessment literacy spans many stakeholders (Jeong, 2013) such as students, teachers, parents, policy-makers, educators and even governments. Itis concerned with certifying learning (Brown, Kennedy,Fok, Chan &

Yu, 2009; Kennedy, Chan & Fok, 2011) and it refers to teachers'theoretical and operational knowledge to design useful tests whose constructs arebased on well-defined test specifications (specs). Further, assessment literacy is interpreted as the teachers' ability to use scores, analyse, comment, report results, make fair inferences on test-takers' future (Qian, 2008) andwrite evaluation reports on instructional materials and language programs(Brown & Bailey, 2008; Davies, 2008a).In this regard, teachers are held to play a key role in assessment reforms.

Acquiring such literacy is imbedded in a broader socio-cultural context where advocates of assessment literacy have addressed assessment conceptions (Brown, 2006; Brown et al. 2009; Gebril & Brown, 2013) and practices (Brown, 2002, 2004, 2008a, 2011). The challenging principle of assessment literacy (Malone, 2013) is to instil a new culture of assessment that straddles conceptions with practices and not disregard them. On the challenges of assessment literacy, Stiggins (2002, p. 762) contends that "few teachers are prepared to face the challenges of classroom assessment because they have not been given the opportunity to learn to do so."Fulcher and Bamford (1996) highlightsome assessment principles that are at the core of any profession where tests are frequently used for examination or promotion. In addition, Inbar-Lourie (2008) stressesthe deployment of multiple principles for assessment literacy, such as the disparity between formative and summative assessment, decision-making, classical vs. modern testing, different modes of assessment, (i.e., alternative and traditional), relevance of formative assessment to instruction, different measurement methods and the societal impacts of assessment practices. In this study, like the original TCoA inventory, the working definition of assessment literacy revolves around four major aspects: *student accountability*, *school accountability*, *improvement* and *irrelevance*.

For Conley (2005), high-stakes assessment is potentially carried out for different purposes the most important of which are improvement and accountability. For instance, teachers use assessment to improve learning and teaching (Brindley, 2001; Scarino, 2013). Hamp-Lyons (1997) argues that the role of the tester is tied with accepting responsibility for all consequences of assessment conceptions and practices. Accountability is one aspect of the hallmark of assessment literacy and it denotes school and teacher responsibility for students' performance. Also, it has the purpose of justifying the worth of money spent on education. New funds are allocated based on students' achievement indicator. Conversely, if accountability is not preserved, assessment is then rendered void, inaccurate and irrelevant. That is, despite the continually cited positive conceptions of assessment, such as student accountability, school accountability and improvement along with its negative washback (Alderson & Wall, 1996), assessment can be viewed as irrelevant (Shohamy, 2001).

Diverse as the assessment purposes might appear, investigating all the different purposes is beyond the scope of this study. To mention few, previous research has cogently investigated TCoA in different contexts, such as New Zealand (Brown, 2006, 2011), China, as an examination-driven context, (Brown & Gao, 2015), Hong

Kong (Brown et al., 2009), the Netherlands (Segers & Tillema, 2011) and Egypt (Gebril &Brown, 2013). Brookhart(2003)and Brown et al. (2009) elevated assessment as a way to improve teaching among New Zealand primary teachers and concluded that teachers had conflicting conceptions of assessment and that they endorsed *Accountability* and *Improvement*. Brown et al. (2009) instrumentally maintained that differences in assessment conceptions were significant, while accentuating the role of assessment in improving and reinforcing learning. Consequently, they envisaged assessment reform in Hong Kong. In Iran,Brown, Pishghadam and Sadafian (2012) underscored *Student Accountability*as being more relevant than *Improvement*. In their study on the effects of high-stakes examination system among pre- and in-service Egyptian teachers, Gebril and Brown (2013) aptly pointed out that the TCoA inventory ofNew Zealand was inadmissible and found a three-factor model, instead of four, with a strong correlation between *Improvement* and *Student Accountability*. Brown (2011) stressed the importance of low-stakes assessment while drawing attention to the significant endorsement between the four main factors of the original TCoA (Brown, 2006).In the observed reviews of the literature on such assessment conceptions, it was found that most of these studies have significantly endorsed some factors at the expense of others whether at the primary, secondary or tertiary level. Nonetheless, no study has investigated TCoA among university and secondary school teachers. This current study addresses this gap in the Tunisian context.

Assessment context in Tunisia

Even though they have their individual, societal, economic,political and educational impingements, test impacts and uses have been overlooked in the Tunisian context. An overwhelming common reform trend induced the furtherance of a change approach by adhering to the implementation of a new educational policy that has been seemingly imbued with a reconsideration of the teaching and assessment practices. Curriculum reform, perceived as tedious, has been open to continuous debate among many stakeholders. While secondary education could be hardly construed by a relentless policy of changing textbooks, tertiary education has bespoken the contrivance of a new educational system labelled as *Bachelor*, *Masters* and *PhD* calling for curriculum reform since 2005. Generally, assessment conceptions, concatenated with the teachers' views of language and language learning, have been marked by different backlogs. There has never been any formal initiative to investigate the Tunisian educational system. There is a dramatic lack in sound theoretical assessment knowledge among teachers and thus, in principle, this lack has most of the time led to poor assessment quality. In this contrived situation, exams are perceived negatively by students, as evidenced by the lack of detailed and adequate assessment reports ontheir performance, apart from a score assigned out of 20. Even in professional events, such as conferences and workshops, there is hardly any organized event on assessment. It is alsoa neglected area at the MA and PhD levels where researchers are not encouraged to investigate the prime considerations, foundations, conceptions and practices of assessment.

Debating this worrying situation has yielded a counterbalancing effect policy seeking refuge in private tutoring that has started to be widespread even at the university level. Given the dramatic absence of governmental stringent guidelines that regulate assessment, a widespread public dissatisfaction with this educational system in general and language assessment in particular has sprung up. It is commonly envisaged that test contents in Tunisia most often answer and meet the teachers' agenda and expectations only, while the role of test-takers to show views of assessment, awareness and even self-assessment has been marginalized. In such an academic milieu, students blame teachers for their wrong conceptions of assessment and their lack of transparency and professionalism and, thus, they revert to absenteeism. Graduates of English "learn" test design out of experience as the scope of education has been flawed in preparing them to be good test designers. Towards such a propagatedassessment policy, students, parents and teachers have largely lost confidence in this system, by labelling assessment as irrelevant. Parents have cautioned against the teachers' policy, and policy-makers have been aspiring for some improvements on the part of teachers.Caught in this vicious circle, no side has claimed responsibility for suchassessment conceptions and practices. Along with these problems, other factors tie in with scoring tests that ispredominantly subjective where specs are not properly delineated to the extent that sometimes test-takers fail to understand the basic requirements of how to process a test item. This is a low-stakes context where there is no plea for the adherence to the implementation of international standardized tests.

The study

The purpose of this study was to investigate secondary and university TCoA and determine the relationship between the four major factors (*School Accountability*, *Student Accountability*, *Improvement* and *Irrelevant*) (Brown, 2006). While studies on TCoA are abundant (Brown, 2011; Gebril & Brown, 2013), there is hardly any study that investigated such conceptions among EFL secondary and university teachers using Brown's inventory of TCoA (2006).In observing all the departments of English at the Tunisian tertiary level up to early 2015, there has been hardly any mention of testing, evaluation and assessment courses (Hidri, 2014), except for an introductory assessment course for the MA students for the year 2012 at the Faculty of Humanities and Social Sciences of Tunis, Tunisia. A major rationale behind this study was to raise assessment awareness among teachers, address the different assessment problems and suggest improvements accordingly. It follows then that the study aimed to answer the following research questions:

1. How do secondary and university teachers conceive of assessment?
2. What does CFA suggest concerning the TCoA inventory held by these teachers?

3. Is TCoA model as undertaken by Brown (2006) similar to the Tunisian TCoA model?

Participants, instruments and procedures

Data of this study were gathered by means of a questionnaire on TCoA inventory (Brown, 2006) administered to university teachers (*n*=206) in 29 institutes and universities and secondary school teachers of English (*n*=336) in the 24 Tunisian governorates. Table 1 presents the demographic features of the participants. Data were collected through a period of nine months, from May 2014 to January 2015. The 336 sample, whose participation rate ranged from .6% to 14.9%, presented 69.9% of females and 30.1% males with 44.6 of teachers having a teaching experience that ranged from 11 to 15 years and .6% for teachers who had a teaching experience of ≥ 30 years with a mean of 3.00. For university teachers,60.2% were females while 39.8% were males with 29.6% of a teaching experience that ranged from 1 to 5 years. The least percentage was among teachers who had an experience of ≥ 30 years with 1.9%. The mean of teaching experience is 2.52 and the participation rate ranged from .5% to 23.3%.

Table 1
Demographic features of the participants (n=542)

Features	Secondary (*n=336*)		University (*n=206*)	
Gender				
	Total	%	Total	%
Female	235	69.9	124	60.2
Male	101	30.1	82	39.8
Teaching Experience				
1-5	48	14.3	61	29.6
6-10	56	16.7	54	26.2
11-15	150	44.6	49	23.8
16-20	24	7.1	20	9.7
21-25	47	14.0	12	5.8
26-30	9	2.7	6	2.9
More than 30	2	.6	4	1.9
Mean		3.00		2.52

The TCoA inventory(Brown, 2006) contains four factors with 27 items or indicators (Appendix A): *School Accountability* and*Student Accountability* as first-order factors, *Improvement*and *Irrelevance,* as second-order factors.*SchoolAccountability* and *Student Accountability* include three first-order contributing indicators each: assessment **provides** information on schools, assessment is **accurate**, assessment

evaluates schools and assessment **categorises** students, assessment **assigns** scores to students' work and assessment **determines**students'qualifications respectively. *Improvement* includes four second-ordercontributing factors, each of which has three other contributing indicators: 1) assessment **describes** abilities (assessment determines the quantity of learning, establishes learning content, and measures meta-cognitive thinking skills among students) 2) assessment**improves learning** (assessment provides feedback on students' performance, feeds back learning needs to students and improves students' learning), 3) **improvesteaching**(assessment is integrated with teaching, modifies teaching and allows different instructions for students) and 4) assessment is **valid**. The fourth factor, *Irrelevance*, has also three second-order contributing factors: 1) assessment is **bad** (against beliefs, unfair, interferes with teaching), 2) assessment is **ignored**(little use of results, results are filed and ignored and impacts teaching) and 3) assessment is **inaccurate**(measurement error, error and imprecision and imprecise process). The scale used in the TCoA inventory was a five-point agreement scale of *strongly disagree, disagree, agree and strongly agree* with a mid-position of"*undecided*". TCoA inventory was administered online to 542 participants. Since the TCoA inventory was administeredonline where the respondents had to select one option from each row to be able to proceed, there were not any missing or invalid cases. Cronbach alpha for the agreement Likert scale wasα= .78 (*M*= 82.86 and *SD*= 13.983) for all the 27 item-data set.

Data Analyses

Given the complexity of factor analysis, data analyses were carried out in four phases (Table 2). The types of analyses were: EFA, parallel analysis(Monte Carlo PCA), dimension analysis and CFAwhich were geared towards defining the appropriate number of factors and checking whether the data fitted Brown's original inventory of TCoA (2006). EFA relied on the use of descriptives of factor extraction, selection and rotation with scree plot and goodness-of-fit-test. Monte Carlo PCA and scree plotstressed the use of eigenvalues, mean and percentile random data eigenvalues, dimension analysis used the SPSS R-Menu v2.0 to get the exact number of factors and CFA was concerned with fit indices, using AMOS v 22.0 so that the data would fit the model.

Phase one:Exploratory Factor Analysis

Using SPSS 22.0, EFA was used as a precursor to CFA to investigate the possible factor structure. In processing EFA (Table 2, phase 1), five statistics were used:

Table 2

Data collection and analyses

Phases	Tests	Results			Rationale
Phase one	*EFA*: - Descriptives: Kaiser-Meyer Olkin (KMO) - Bartlett's Test of Sphericity - Factor extraction (ML), - Factor selection (Eigenvalues and scree plot, and goodness-of-fit-test - Factor rotation (Promax)	KMO: .794 χ^2: 2487.726 *df*: 351 *p*= .000 8 factors χ^2: 276.832, *df*: 163, *p*= .000	KMO: .782 χ^2: 2487.726 *df*: 253 *p*= .000 6 factors χ^2: 226.130, *df*: 130, *p*= .000		- To get the possible factor structure - To define the items whose values are ≥ .30 - To check if the ratio of χ^2/*df*) are significant at *p*= .000
Phase two	- *Parallel analysis:*Monte Carlo PCA - Scree plot	6 factors 6 factors			- To check the number of factors that had to fit the model.
Phase three	- *Dimension analysis*: SPSS R-Menu v.2.0. - RMSR eigenvalues - Parallel analysis - Goodness-of-fit-test	3 to 8 factors Eigenvalues>mean = (*n*= 8), parallel analysis (*n*= *6*), optimal coordinates (*n*= *3*) χ^2: 1027.58, *df*: 273, *p*= .000			- To check the number of factors yielded by parallel analysis
Phase four	*CFA*: AMOS 21.0	Original model (Appendix B, 27 items)	6 factors (Appendix C, 22 items)	3 factors (TunisianTCoA (Figure 5)(15 items)	- To check if the fit value is acceptable
	- χ^2, *df*, χ^2/*df*, SRMR, RMSEA, CFI and TLI Reliability analysis: Cronbach Alpha Pearson Correlation analysis	- χ^2: 1805.788 - *df*: 312 - $\chi^{2/df}$: 5.788 - SRMR: .17 - RMSEA: .094 - CFI: .52 - TLI: .46 - *p*= .000 - *α*= .78 (*M*= 82.86, *SD*= 13.983)	- χ^2: 525.278 - *df*: 194 - $\chi^{2/df}$: 2.708 - SRMR: .10 - RMSEA: .056 - CFI: .84 - TLI: .80 - *p*= .000 - *α*= .75 (*M*= 70.15, *SD*= 12.087)	- χ^2: 266.872 - *df*: 87 - $\chi^{2/df}$: 3.067 - SRMR: .089 - RMSEA: .062 - CFI: .90 - TLI: .88 - *p*= .000 - *α*= .78 (*M*= 56.93, *SD*= 9.68)	- To check the goodness-of-fit indices (since relying on χ^2might not produce valid fit indices).

descriptives (KMO), Bartlett's Test of Sphericity, factor extraction (Maximum Likelihood (ML)), factor selection (eigenvalues and scree plot) and factor rotation loadings (goodness-of-fit-test and Promax). Factor extraction and factor rotation using ML were meant to define the exact number of factor loadings to fit the model. To produce high loadings whose values should be ≥ .30, the pattern matrix was considered to investigate the loadings that were the result of the appropriate relationship between factors and indicators. Therefore, the indicators that were ≤ .30 or the ones which were linked to more than one factor were deselected.

Phase two: Parallel analysis

After the rotation phase, a parallel analysis was used (see Hayton, Allen, & Scarpello, 2004 on parallel analysis). Specifically, the Monte Carlo PCA parallel analysis (e.g., scree plot) was carried out to determine the statistically significant eigenvalues based on random data generation. The sample data was set at 100 cases to produce a valid analysis. The desired percentile was set at 95%, then a PCA was set at 2 and the data generation parallel analysis was set at 1. These tests were implemented to determine the factor structure.

Phase three: Dimension analysis

Dimension analysis of the data using the SPSS R-Menu v2.0 was conducted to opt for"statistical and graphical computing" (Courtney, 2013, p. 5) to estimate and reach the right number of factors. Unlike EFA, dimension analysis produced different results from the ones of phases one and two.

Phase four: Confirmatory factor analysis

Extant research has shown that CFA,pertained to Structural Equating Model, has been widely used to investigate the relationship paths between variables,the correlation between observed and latent variables (Miller, Davidson, Schindler, & Messier, 2013), to test data fit(Brigman, Wells, Webb, Villares, Carey &Harrington, 2015)and to check indicators' influence on factors(Hu & Bentler, 1998, 1999). CFA was carried out using AMOS to check whether the data would yield another Tunisian inventory, since the original model was found to be inadmissible. Many researchers have maintained that a sample of ≤ 400 respondents is appropriate to claim validity of CFA results (Brown, 2006). CFA includes the following criteria: chi-square statistic (χ^2), degree of freedom (*df*) and ratio of chi-square to degrees of freedom ($\chi^{2/df}$). To counterbalance chi-square sensitivity, four indices were considered: Root mean squared error of approximation (RMSEA), root mean square residual (RMSR), Tucker–Lewis non-normed fit index (TLI) and comparative fit index (CFI).

Values of CFA range from .0 to 1 (Brown, 2006; Cokley, 2014).In addressing the latent trait modeling, Brown (2006), Browne and Cudeck (1992) and Hu and Bentler (1999) suggested the following good fit indices values: CFI ≤.95, RMSEA ≈.06. For other researchers (e.g., Hair, Tatham, Anderson & Black, 2005), the RMSEA value of0.10 is indicative of unacceptable fit, 0.08-0.10, mediocre fit, 0.06-0.08, acceptable fit, and 0.01-0.06, close fit. In all cases, it should be<.70. The$\chi^{2/df}$ ratio should range from 2 to 3as indicative of good fit (Choo, Walsh & Teyl, 2013); while values of CFI and TLI should be 0.95 as good fit and 0.90 as acceptable fit. For SRMR, values of 0.08 indicate good fit and 0.12 would indicate acceptable fit. The goodness-of-fit indices were considered based on the suggestions of Brown (2006) and Hu and Bentler (1998, 1999).

Results

EFA results

All the 27 items of the original TCoA were considered in order to yield a possible factor structure of the appropriate latent variables and indicators with high loadings, and, therefore, check whether this inventory was admissible to the Tunisian context.Like the original model (Brown, 2006), six first order items (Appendix B) were loaded onto*School Accountability* and *Student Accountability* respectively. Four second-order items with three items each were loaded onto*Improvement*. As for the fourth factor (*Irrelevance*), three items included three contributing factors each respectively.At this level, the model was found to be inadmissible. Therefore, EFA was carried out to define the exact number of factors to yield appropriate fit indices.For instance, EFA using parallel analysis, pattern matrix and scree tests suggested an eight-factor model. Loadings of ≤ .30 were discarded.The rotated factors using ML were estimated at an eigenvalues of ≥.30. The *KMO* was .794 indicating an acceptable fit value of all the data (Table 2). The Bartlett's test of sphericity $\chi^2$was 2487.726 with

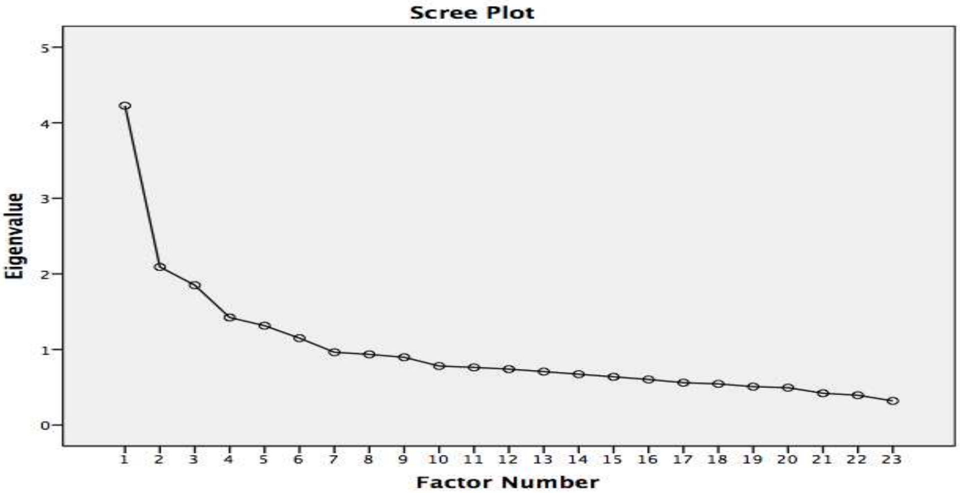


*Figure 1.*Scree plot of the eigenvalues (EFA, 8 factors)

a *df* of 351. The goodness-of-fit-test of $\chi^2$was 276.832 with a *df* of 163was significant at .000, as expected because of the huge dataset (*n=542*). Based on the eigenvalues of the 27 indicators, Figure 1 shows 8 factors that the data set produced. The number of factors was determined once the slope of the figure changed. The number of factors was determined by counting all the instances that occurred before this slope. However, Brown (2006) claims that the results of scree test may sometimes be ambiguous.

Table 3

Raw data eigenvalues, mean and percentile random data eigenvalues

Root	Raw Data	Means	Prcntyle
1.000000	4.764549	1.436147	1.491098
2.000000	2.370154	1.372303	1.424023
3.000000	2.006776	1.324681	1.363655
4.000000	1.526081	1.282012	1.320723
5.000000	1.398338	1.241457	1.270199
6.000000	1.265056	1.206071	1.231711
7.000000	1.148393	1.174377	1.196837
8.000000	1.102225	1.146129	1.168262
9.000000	.990467	1.117137	1.145560
10.000000	.919678	1.089348	1.111021

Analysis of the pattern matrix yielded unfit values in the dataset at the initial phase with some outliers. To remedy this, the items that did not load anywhere or the ones that loaded with two factors were deselected from the pattern matrix analysis, such as items 2, 3, 10, 12 and 22. Therefore, the pattern matrix yielded loadings ranging from .334 (indicator 9) to .935 (indicator 20). The *KMO* measure of sampling adequacy was .782 and χ^2 was 2487.726 with a *df* of 253 was still significant at .000. The goodness-of-fit-test for the chi-square was 226.130 with *df* of 130, significant at .000. EFA results indicated the loadings of six factors along withtheir commonalities: E.g., 3 (.756), 4 (.703), 14 (.632), 5 (.452), 13 (.431) and 1 (.423) loaded on factor 1; indicators 20 (.935), 21 (.656), and 19 (.576) loaded on factor 2, indicators 27 (.607), 11 (.547), 6 (.506), 16 (.472) and 15 (.412) loaded on factor 3, indicators 24 (.683), 25 (.549) and 23 (.382) loaded on factor 4, indicators 17 (.659), 18 (.451) and 26 (.357) loaded on factor 5 and indicators 8 (.600), 7 (.497) and 9 (.334) loaded on factor 6. Still, at this level, the factors had conflicting indicators. To solve this problem, parallel analysis was implemented.

Parallel results

Parallel analysis was utilized to further check the number of factors that had to fit the model. Table 3indicates results of the Monte Carlo PCA analysis that yielded six factors. It presents raw data eigenvalues, mean and percentile random data eigenvalues carried out onthe 27 items for all the participants (N= 542), 10 items of the framework and 100 random data sets were generated and a percentile of 95. Column two, raw data, calculated the PCA of eigenvalues on the correlation matrix that matched the actual data of the SPSS. The first item has a component eigenvalue of 4.764549, the mean (which was anchored at 50%) is 1.436147, and a percentile value of 1.491098 anchored at 95 up to item 6 with an eigenvalue of 1.265056, mean of 1.206071 and a percentile of 1.231711. The percentile has to be larger than the mean of eigenvalues and it, therefore, indicates a factor. The eigenvalues of thefirstsix raw data values were statistically significant, since their values were larger than the benchmark criterion value of the percentile. Based on this analysis, data analysis

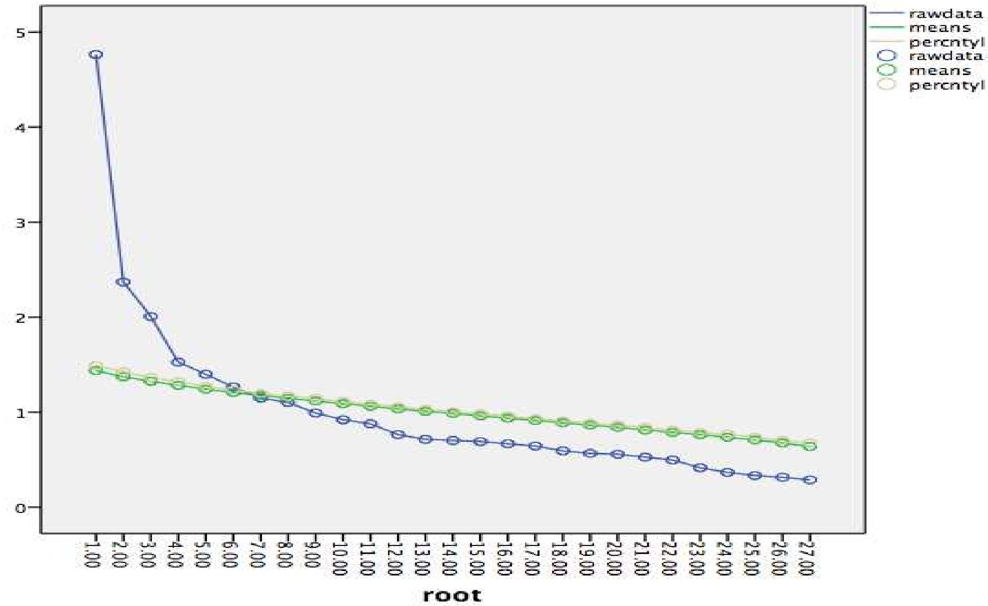


Figure 2. Data sequence plot (Monte Carlo PCA)

pointed to six factor components. A graphical presentation of the Monte Carlo PCA using scree plot (Figure 2) shows six factors. The (blue) line with the vertical orientation presented the raw data, the green line indicated the means and the last one (brown line) showedthe eigenvaluesand it, therefore, showed a factor. The eigenvalues of thefirst six raw data values were statistically significant, since their values were larger than the benchmark criterion value of the percentile. Based on this analysis, data of the study pointed to six factor components to extract from this analysis. The two lines (brown and green) intersected with the scree, vertical line, and factors above the competing eigenvalue lines represented the number of factors that should be extracted. The number of factors was selected based on the intersection lines. Therefore, the points above the intersection lines represented the number of factors, which showed 6 factors.

Dimension analysis results

In the dimension analysis results, a correlation matrix of heterogeneous (two steps) analysis indicated the retention of 3 factors (Figure 3). However, the analysis of comparison data (fit-to-comparison data and Pearson analysis), (Figure 4), indicated the retention of 8 factors. Results of dimension analyses using the SPSS R-Menu v2.0

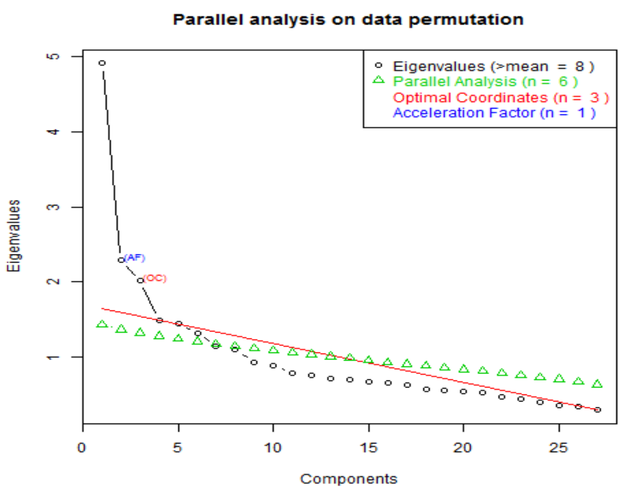


Figure 3. Parallel analysis on data permutation

Table 4
Factor correlation matrix

Factor	1	2	3
1	1.000	.187	-.240
2	.187	1.000	-.118
3	-.240	-.118	1.000

Extraction Method: Maximum Likelihood.
Rotation Method: Oblimin with Kaiser Normalization.

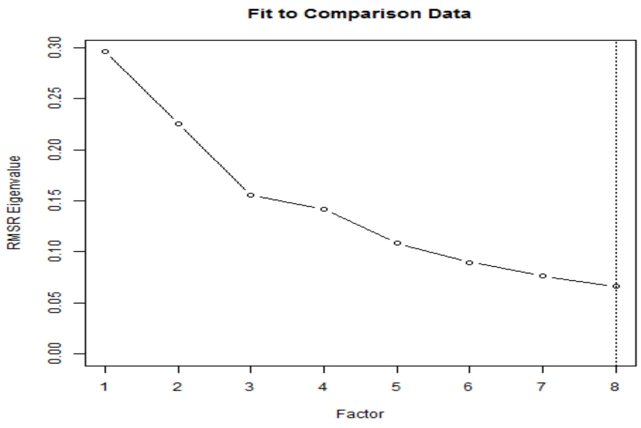


Figure 4. Fit to comparison data (8 factors)

Table 5
Fit to comparison dataof 8 factors

Nb. of factors	RMSR Eigenvalue	p-value
1 factor	.296	NA
2 factor	.225	.000
3 factor	.156	.000
4 factor	.142	.000
5 factor	.108	.000
6 factor	.090	.000
7 factor	.076	.000
8 factor	.066	.000

(Courtney, 2013) showed a data set that ranged from 3 (Figure 3) to 8 factors (Figure 4) and the factor correlation matrix (Table 4) showed 3 factors, however, the fit to comparison data eigenvalues indicated 8 factors (Table 4).

CFA results

Once the number of factors was determined, CFA was used. Data of the study on the TCoA in Tunisia indicated a range of 3 to 8 factors.The initial phase of the study constituted of the use of CFA. Recall that AMOS covariance matrices analyses between factors demonstrated that Brown's model (2006) was found to be inadmissible (Appendix B) with the following fit indices: χ^2 = 1805.788, *df*= 312, $\chi^{2/df}$ = 5.788, CFI= .52, TLI= .46, and RMSEA= .094 and SRMR= .17. The loadings from *Improvement* to *describe* and *improves learning* were beyond the range as they both had a value of 1.15 and 1.02 and 1.39 respectively. Recall that in CFA, such values should range from .0 to 1. Also, the covariance matrices between *Improvement* and *Student Accountability* on the one hand and between *School Accountability* and *Irrelevance* on the other were beyond the range with values of 1.15 and 1.53 respectively.

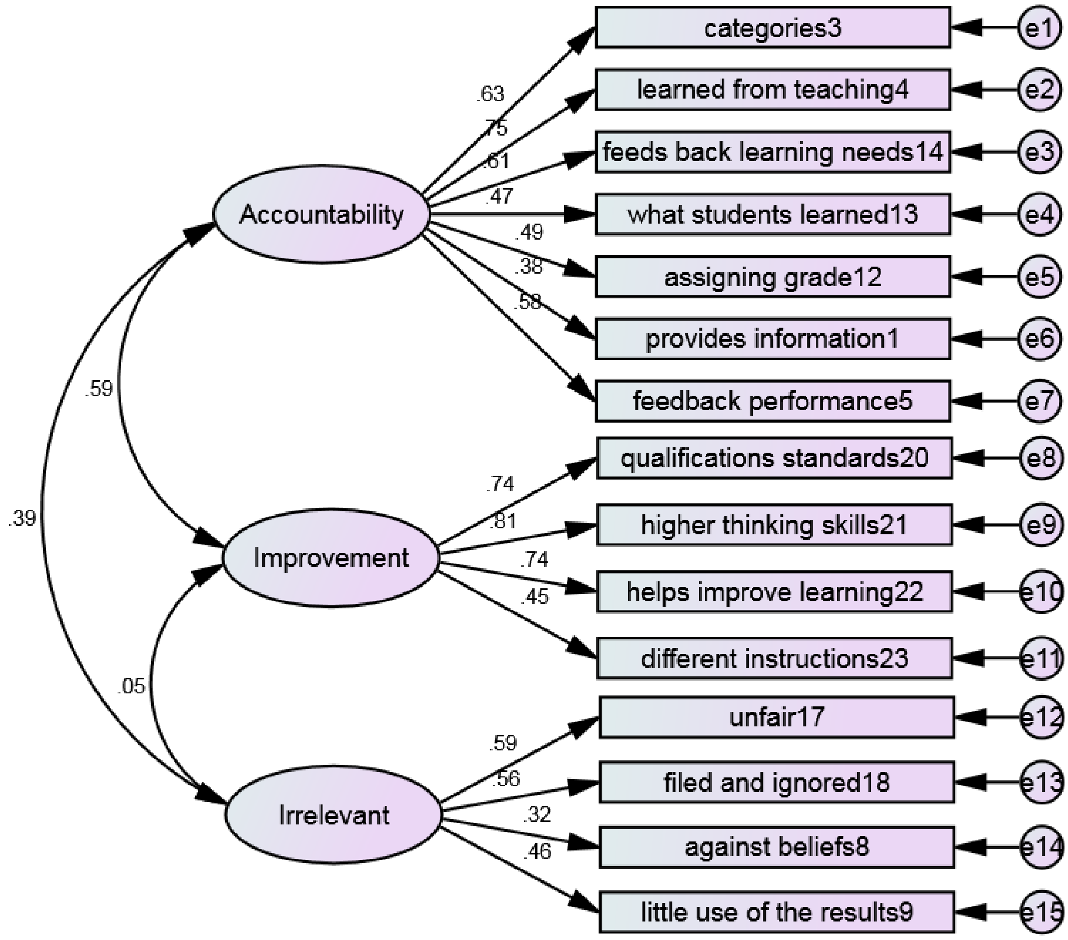


Figure 5. The Tunisian model of TCoA:χ^2: 266.872, *df*: 87. $\chi^{2/df}$: 3.067, SRMR: .89, CFI: .90, TLI: .88, RMSEA: 062. All factor loadings were significant at *p*= .000 level.

After deselecting indicators 2, 3, 10, 12 and 22, the model yielded 6 factors with acceptable fit indices (see Appendix C): *p*= .000, χ^2: 525.278, *df*: 194. $\chi^{2/df}$: 2.708, SRMR: .10, CFI: .84, TLI: .80, RMSEA: .056. However, even though the model had high loadings, it nonetheless was inadmissible, since many indicators did not load onto the appropriate factors, such as 19, 27, 26 and 7. In trying to attach the indicators with the appropriate factors, the preliminary Tunisian model (Appendix C) was edited and the deselected factors were reintegrated in CFA to recheck the factor structure. Figure 5 presents the following three-factor model with first order factors attached to 15 indicators. Factor 1, *Accountability*, included indicators and loadings as follows: 3 (*r*= .63), 4 (*r*= .75), 14 (*r*= .61), 13 (*r*= .47), 12 (*r*= .49), 1 (*r*= .38) and 5 (*r*= 58). Factor 2, *Improvement*, included indicators 20 (*r*= .74), 21 (*r*= .81), 22 (*r*= .74) and 23 (*r*= .45) and factor 3, *Irrelevant*, included indicators 17 (*r*= .59), 18 (*r*= .56), 8 (*r*= .32) and 9 (*r*= .46). The fit indices of the Tunisian model are the following: p= .000, χ^2: 266.872, *df*: 87. $\chi^{2/df}$: 3.067, SRMR: .89, CFI: .90, TLI: .88, RMSEA: 062 and, therefore, a three-factor trimmed model version was found to be admissible. It could be cautiously argued that this model somehow yielded acceptable fit indices.

In addressing the internal consistency of the factors, the reliability coefficients indices using Cronbach Alpha (α) were considered for both populations: secondary school and university teachers. Reliability of the three factors was calculated among both populations. Results of*Accountability* among secondary school teachers indicated α= .81 (*M*= 23.05, *SD*= 5.984), *Improvement* α= .78 (*M*= 12.96, *SD*= 3.738) and *Irrelevant* α= .63 (*M*= 12.60, *SD*= 3.730). As for university teachers data, *Accountability*α= .60 (*M*= 20.67, *SD*= 5.045), *Improvement* α= .76 (*M*= 13.16, *SD*= 3.847) and *Irrelevant* α= .16 (*M*= 10.33, *SD*= 2.907). The entire scale for both populations yielded the following results: α= .75 for *Accountability* of 7 items (*M*=22.15, *SD*= 5.758), α= .77 for *Improvement* of 4 items (*M*= 13.04, *SD*= 3.778), α= .54 for *Irrelevant* of 4 items (*M*=11.74, *SD*= 3.610) and α= .78 for the entire scale statistics of 15 items (*M*= 56.93, *SD*= 9.68).

In addition, a Pearson correlation coefficient analysis was conducted to investigate the correlation patterns between the 15 indicators. For results on secondary school teachers' data, the *Accountability*correlation between*learned from teaching* and *feedback performance,* with coefficients of .55,was significant at (*p*≤. 001). The least significant correlation between*provides information* and *feedback performance*was .28 (*p*≤. 001). For*Improvement*, correlation coefficients between *higher thinking skills* and *qualification standards*were .63 (*p*≤. 001), while they were significant at .31 (*p*≤. 001) between *different instructions* and *higher thinking skills.* For*Irrelevant*, correlation coefficients between *little use of results* and *against beliefs*were .41 (*p*≤. 001), however they were .15 (*p*≤. 001) between *against beliefs*and *filed and ignored*.For *Accountability* results among university teachers, the Pearson correlation coefficients between *categories* and *learned from teaching* were .40 (*p*≤. 005) and .19 (*p*≤. 005) between *what students learned* and *feeds back learning needs,* while the coefficients were .14 (*p*≤. 001) between *assigning a grade* and *categories*. As for *Improvement*, correlation coefficients between *higher thinking skills* and *qualifications standards*were .62 (*p* ≤. 001) and .25 (*p* ≤. 001) between *higher thinking skills* and *different instructions*. As for *Irrelevant,* correlation coefficients between *unfair* and *filed and ignored*were .33 (*p* ≤. 001). However, the Pearson correlation coefficientswere not significant since they yielded the following: -.072 between *filed and ignored* and *against beliefs*. In observing the model, the *Accountability*correlation between *learned from teaching* and *categories*was .49 ($p \leq .001$), while the correlation between *feeds back learning needs* and *providesinformation* was .17 ($p \leq .001$). As for *Improvement*, the Pearson correlation coefficients between *higher thinking skills* and *qualifications standards*were .63 ($p \leq .001$), butwere .27 ($p \leq .001$) between *different instructions* and *higher thinking skills*.As for *Irrelevant*, the Pearson correlation coefficients between *unfair* and *filedand ignored*were .36 ($p \leq .005$), while they were .20 ($p \leq .005$) between *unfair* and *against beliefs*werebut.09 ($p \leq .001$) between *filed and ignored* and *against beliefs*.

Discussion

The study investigated secondary and university TCoA in an EFL context. To approach this, different analyses were adopted. A conspicuous result of the study was levelled at the wrong and conflicting assessment conceptions among Tunisian teachers of English. The latent structural paths between variables and indicators of the initial model were different from previously observed models (e.g., Brown, 2006; Brown & Goa, 2015; Gebril & Brown, 2013) and, therefore, it resulted in poor fit. This is reflected in the beyond-range valuesof relationship (See Appendix B for covariance matrices). However, based on the different types of analyses, one may provide a baseline that the data collected on Tunisian secondary and university teachers partially fitted the model, resulting in high loadings, but with different configurations of the relationship between factors and indicators. This partial fit indicated that assessment conceptions were held to be divergent among teachers of English.

While previous studies highlighted the results of 27 items with appropriate loadings but with different combinations of factors and indicators (e.g., Brown & Michaelides, 2011), this study had different results in that the 15 indicators loaded onto different factors. Structurally speaking, the paths among factors and between factors and indicators (Appendix C) were disparate and divergent from the original TCoA (indicators 19, 27, 26 and 7 are a case in point) even though they yielded high loadings. However, like other studies (e.g., Kitiashivili, 2014), a central axiom here lies at the heart of the conjointly problematic assessment misconceptions among secondary and university teachers despite the correlation of high loadings. This had to be expected given the stakeholders' mundane attitudes towards assessment where assessment courses or professional development events are not attributed their due relevance. Thus, this lack of assessment expertise does not bode well for a clear and sustained assessment policy. An essential comment of this discussion that can be upheld is that considering relevant assessment literacy need not be ignored, nor need it be taken for granted. Rather, it should be based on objective and well-sustained guidelines where teachers are expected to hold a key role. The conflicting conceptions of assessment might also impact teachers' practices. This idea was echoed in the study conducted by Cheng, Rogers and Hu(2004).

The Tunisian model of TCoA was divergent from previous studies' models, such as Brown and Michaelides(2011) and Brown(2011). Even though the Tunisian model might look similar to the simplified New Zealand model, it is nevertheless different, since it resulted in a trimmed number of factors and indicators, 15 instead of 27 (see Appendix D on the number of deselected items from the original TCoA). The Tunisian model of TCoA had direct relationships between indicators and their factors. Like previous studies (e.g., Gebril & Brown, 2013), this study showed a strong correlation between *Accountability* and *Improvement*. What could be deduced is that despite the fact that assessment has not been attributed its role in such context, teachers still conceive of assessment in a positive way. This result is echoed in other studies (e.g., Brown, 2006). It could also be concluded that secondary and university TCoA wereenmeshed in disparate attitudes in a sense that whileteachers conceived of assessment as a way to improve learning, they, also, perceived it as irrelevant. In the Tunisian context, teachers have been held responsible for the testing quality on the ground that teaching and testing do not prepare graduates to be operational in their field of work. In this regard, students' criticism has been lodged at teachers whose exams have been perceived as a heavy burden. Meaning-ladenness of assessment has been suffused with vagueness and disparity given the dramatic lacks in theoretical underpinnings and practical tips of assessment and unfortunately, students graduate with no basic knowledge or training in assessment. Notwithstanding the *ad hoc* leaps in reform that have marked education in Tunisia, still, the assessment policy is vague, as teachers who do not have clear assessment visions are likely to miss the learning objectives and, therefore, make of testing an irrelevant task. Other stakeholders such as learners, parents and policy-makers should attempt to take cognizance of the teachers' divergent educational backgrounds. This is manifested in the wrong loadings of indicators onto the latent variables, which could be truly due to the misspecifications of the TCoA.

Unlike the previous educational system that rested on summative assessment, newer developments within the current assessment policy in Tunisiaproffered the idea of a revamp offormative assessment. Unfortunately, unlike other studies (e.g., Brindley, 2001),assessment in Tunisia has been perceived as a way to evaluate students on whether to pass or fail, and not as a way to develop their critical thinking skills to overcome the different language problem-solving tasks that tests may contain. To strive against the teachers' bias for summative assessment, many stakeholders such as policy-makers, should consider the implementation of both formative and summative assessment. Whether at school or university, one result of this study showed that assessment was generally used in a summative way, and not as a diagnostic tool to establish a comprehensive view about the students' language ability. Item writers overlook test specs on the ground that they are experienced enough to design fair tests. One trenchant criticism at this level is thattest designers write items that measure a limited range of skills and sub-skills or tests that do not measure what they are intended to measure.This reflected the teachers' lack of sound knowledge about language assessment.

Limitations, implicationsand recommendations

Results of the study indicated that much research is needed to investigate the TCoA at the different educational levels in Tunisia. Analysing data of the study on secondary and university teachers separately could have led to other results. This posed some limitations, since the implementation of factor analysis necessitates the use of ≥ 400 participants. Investigating comparison of the model between the two populations might enlighten the researcher to compare and contrast assessment conceptions in the two models. In addition, using research mixed methods, such as interviews and item analysis of test scores, could undoubtedly provide more insights into the teachers' assessment conceptions that would have direct implications for teaching as well as assessment.

In addressing implications of the study, it could be argued that, as part of methodological implications, the use of factor analysis (EFA, PCA, SPSS R v.2.0 and CFA) helped to unveilassessment conceptions and toaddress the factor structure of acceptable or good fit models. In the pedagogical implications, this study presented practical steps in how to investigate teachers' beliefs about assessment, which are directly linked toclassroom practices. Like other studies (Gebril & Brown, 2013) the current study might plausibly be relevant to such contexts in probing into the major requirements of assessment as well as teaching conceptions. As for the research implications, addressing such assessment enterprises was intertwined with some wrong assessmentbeliefs and views thatcan potentially affect assessment practices. Such practices are upheld by many teachers who are encouraged to invest a lot in formative assessment to provide test-takers with continuous feedback on their performance. Further research on the teachers' conceptions of assessment is needed in the Tunisian context. An exhortation to teachers is to team up, consider test specs, pilot test items, administer tests, analyse test scores and reconsider test specs where collaborative initiativesshould be highlighted to improve the assessment quality. In line with the different calls for change, all stakeholders should seize this opportunity and consider all challenges pertained to assessment. It is highly recommended that universities in Tunisia reconsider their curricula to stress the relevance of administering assessment courses for students before graduation.

One of the outcomes of the study was the call for MA and PhD students in particular and researchers in general to re-investigate the teachers' assessment conceptions and by extension practices.The adherence to a code of practice among all stakeholdersis a diligent necessity. In addition, implementing high-stakes professional assessment standards should be pertained to the learners' knowledge of instructional objectives and testing outcomes. In this regard, test designers should be aware of the different testing methods, such as classical, modern, discrete-point, integrative or communicative, since this awareness reflects their views of language and language learning. Another challenge that faces test designers is the dilemma of whether the training in testing, if any, can enable them to design useful tests.To develop their assessment literacy, teachers should consider the following: test specs, assessment ethics, professional assessment standards, testing outcomes, relevance of placement tests, alternative forms of assessment, setting up exam boards, needs analysis of students' assessment lacks, relevance of using international standardized exams as entry and/ or exit exams, item analysis, testing and curriculum design, critical approaches to assessment, program evaluation and classroom testing practices. Since assessment is the backbone of any educational system, assessment courses should be considered at the university level. As for secondary school teachers, supervisors of English themselves have to target assessment literacy in their regular training sessions. Whether at the secondary or university level, assessment in Tunisia has to be taken seriously and meaningfully. However, itis still norm-referenced and it has been widely implemented at the expense of criterion-referenced assessment and probably teachers, test designers and other stakeholders have not been aware of this dichotomy.

Acknowledgements

The researcher would like to express his deep gratitude to Prof. Gavin T L Brown for his invaluable comments and suggestions to reach a final shape of TCoA in Tunisia.
A special thank-you would also go to the participants who volunteered to take part in the study and to the people who helped with data collection. Data interpretations are those of the researcher and they do not reflect the participants' ideas.

References

Alderson, J. C., & Wall, D. (Eds.). (1996). Special issue on washback. *Language Testing*, 13(3).

Brigman, G., Wells, C., Webb, L., Vallares. E., Carey. C. J., &Harrigton, K. (2015). Psychometric properties and confirmatory factor analysis of the student engagement in school success skills. *Measurement and Evaluation in Counseling and Development*, 48(1) 3-14. doi: 10.1177/0748175614544545

Brindley, G. (2001). Outcomes-based assessment in practice: some examples and emerging insights. *Language Testing,* 18 (4) 393-407. doi: 02655322(01)LT214OA

Brookhart, S. M. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues and practices*, 22(4), 5-12.

Brown, G. T. L. (2002). Conceptions of assessment III abridged survey. Research Survey Published by the University of Auckland, NZ. Retrieved July 26, 2007, from http://www.arts.auckland.ac.nz/FileGet.cfm?ID=0839d599-6870-480a-b922-40ccb60fa1e5

Brown, G. T. L. (2004). Teachers' conceptions of assessment: Implications for policy and professional development. *Assessment in Education: Principles, Policy & Practice*, 11(3), 301-318.

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

Brown, D. J. &Bailey, K. M. (2008). Language testing courses: What are they in 2007? *Language Testing*, 25(3) 349-38.

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

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, 347-363.

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. & Michaelides, M. (2011). Ecological rationality in teachers' conceptions of assessment across samples from Cyprus and New Zealand. *European Journal of Psychological Education*, 26: 319-337.

Brown, G. T. L., Pishghadam, R., & Sadafian, S. S. (2014). Iranian university students' conceptions of assessment: Using assessment to self-improve. *Assessment Matters*, 2014, 6 pp. 5-33.

Brown, G. T. L & Gao, L. (2015). Chinese teachers' conceptions of assessment for and of learning: Six competing and complementary purposes. *Cogent Education*, 2: 993836, 1-19. http://dx.doi.org/10.1080/2331186X.2014.993836

Brown, T. A. (Ed.) (2006). *Confirmatory factor analysis for applied research*. The Guilford Press.

Browne, M. W. & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods and Research,* 21: 230-258.

Cheng, L., Rogers, T., & Hu. H. (2004) ESL/EFL instructors' classroom assessment practices: purposes, methods, and procedures. *Language Testing,* 21(3) 360-389. doi: 10.1191/0265532204lt288oa

Choo, W. Y., Walsh, K. C. & Teyl, N. P. (2013). Teacher Reporting Attitudes Scale (TRAS): Confirmatory and Exploratory Factor Analyses With a Malaysian Sample. *Journal of Interpersonal Violence,* 28(2) 231-253. doi: 10.1177/0886260512454720

Cokley, K. (2014). A Confirmatory Factor Analysis of the Academic Motivation Scale With Black College Students. *Measurement and Evaluation in Counseling and Development,* 1-16. doi: 10.1177/0748175614563316

Conley, M. W. (2005). *Connecting standards and assessment through literacy*. Boston: Allyn & Bacon.

Courtney, M. G. R. (2013). Determining the Number of Factors to Retain in EFA: Using the SPSS R-Menu v2.0 to Make More Judicious Estimations. *Practical Assessment, Research and Evaluation*, 1-14.

Davies, A. (2008a). Textbook trends in teaching language testing. *Language Testing,* 25(3), 327-347.

Fulcher, G & Bamford, R. (1996). I didn't get the grade I need. Where's my solicitor?

System, 24, No. 4, pp. 437-448.

Gebril, A. & Brown, G. T.L. (2013). The effect of high-stakes examination system on teacher beliefs: Egyptian teachers' conceptions of assessment. *Assessment in Education: Principles, Policy and Practice*, 21, No 1, 16-33. doi: 10.1080/0969594X.2013.831030

Hair, J. F., Tatham, R. L., Anderson, R. E., & Black, W. C. (2005). *Multivariate data analysis* (6th Ed.). New York, NY: Prentice Hall.

Hamp-Lyons, L. 1997: Washback, impact and validity: ethical concerns. *Language Testing,* 14: 295-303.

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*, 7, 191-205. doi: 10.1177/1094428104263675

Hidri, S. (2014). Developing and evaluating a dynamic assessment of listening comprehension in an EFL context. *Language Testing in Asia, 4:4.* doi: 10.1186/2229-0443-4-4

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

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to under-parameterization model misspecification. *Psychological Methods, 3*, 424-453. doi: 10.1037/1082-989X.3.4.424

Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing,* 25(3) 385-402.

Jeong, H. (2013). Defining assessment literacy: Is it different for language testers and non-language testers? *Language Testing,* 30(3) 345-362. doi: 10.1177/0265532213480334

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, 41–54. doi:10.1080/14748460.2011.550433

Kitiashvili, A. (2014). Teachers' attitudes toward assessment of student learning and teacher assessment practices in general educational institutions: The case of Georgia. *Improving Schools*, 17(2) 163-175. doi: 10.1177/1365480214534543

Lado, R. (1961). *Language testing*. London: Longman.

Malone. M. (2013). The essentials of assessment literacy: Contrasts between testers and users. *Language Testing,* 30(3) 329-344. doi: 10.1177/0265532213480129

Miller, D., Davidson, P., R., P., Schindler, D., & Messier, C., (2013). Confirmatory Factor Analysis of the WAIS-IV and WMS-IV in Older Adults, *Journal of Psychoeducational Assessment,* 31(4) 375-390. doi: 10.1177/0734282912467961

Qian, D. D., (2008) English language assessment in Hong Kong: A survey of practices, developments and issues. *Language Testing,* 25(1) 85-110. doi: 10.1177/0265532207083746

Scarino, A. (2013). Language assessment literacy as self-awareness: *Understanding* the role of interpretation in assessment and in teacher learning. *Language Testing,* 30(3) 309-327. doi: 10.1177/0265532213480128

Segers, M, & Tillema, H. (2011). How do Dutch secondary teachers and students perceive the purpose of assessment? *Studies in Educational Education (special issue)*, 37, 49-54.

Shohamy, E. (2001). Democratic assessment as an alternative. *Language Testing, 18(4), 373-391.*

Stiggins, R. J. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan*, 83(10), 758-765. Retrieved October 9, 2014, from http://pdk.sagepub.com/content/83/10/758.full.pdf+html

Stoynoff, S., & Chapelle, C. (2005). ESOL tests and testing: A resource for teachers and program administrators. Alexandria, VA: TESOL Publications.

Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics*, 29, 21-36.

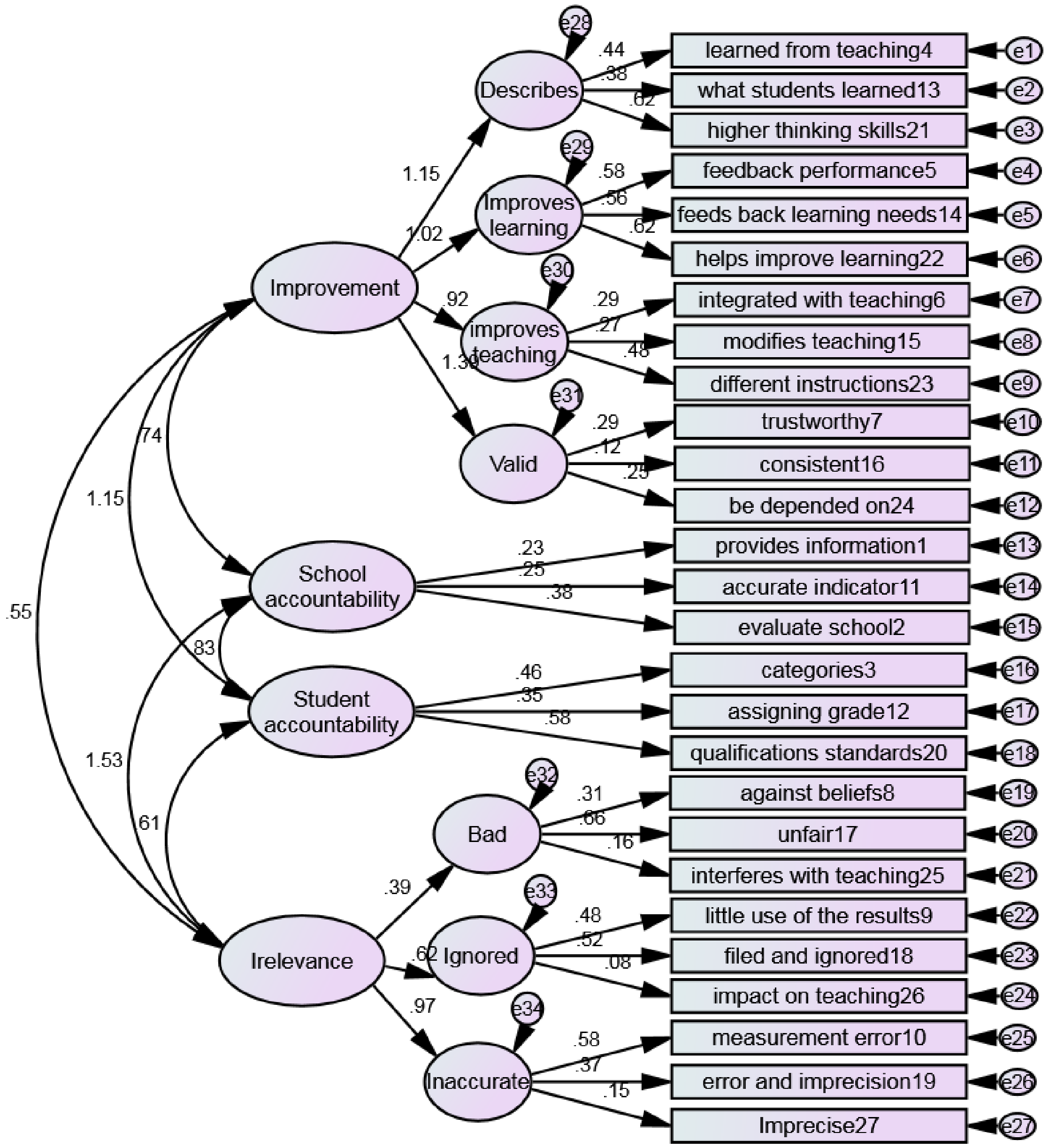
Appendix A Teachers' Conceptions of Assessment (Brown, 2006, p. 168)

Factors and indicators

- Assessment Makes Schools Accountable
  - Assessment provides information on how well schools are doing.
  - Assessment is an accurate indicator of a school's quality.
  - Assessment is a good way to evaluate a school.
- Assessment Makes Students Accountable
  - Assessment places students into categories. .
  - Assessment is assigning a grade or level to student work.
  - Assessment determines if students meet qualifications standards.
- Assessment Improves Education
  - Assessment Describes Abilities
    - Assessment is a way to determine how much students have learned from teaching.
    - Assessment establishes what students have learned.
    - Assessment measures students' higher order thinking skills.
  - Assessment Improves Learning
    - Assessment provides feedback to students about their performance.
    - Assessment feeds back to students their learning needs.
    - Assessment helps students improve their learning.
  - Assessment Improves Teaching
    - Assessment is integrated with teaching practice.
    - Assessment information modifies ongoing teaching of students.
    - Assessment allows different students to get different instruction.
  - Assessment is Valid.
    - Assessment results are trustworthy.
    - Assessment results are consistent.
    - Assessment results can be depended on.
- Assessment is Irrelevant
  - Assessment is Bad
    - Assessment forces teachers to teach in a way against their beliefs.
    - Assessment is unfair to students.
    - Assessment interferes with teaching.
  - Assessment is Ignored
    - Teachers conduct assessments but make little use of the results.
    - Assessment results are filed and ignored.
    - Assessment has little impact on teaching.
  - Assessment is Inaccurate
    - Assessment results should be treated cautiously given measurement error.

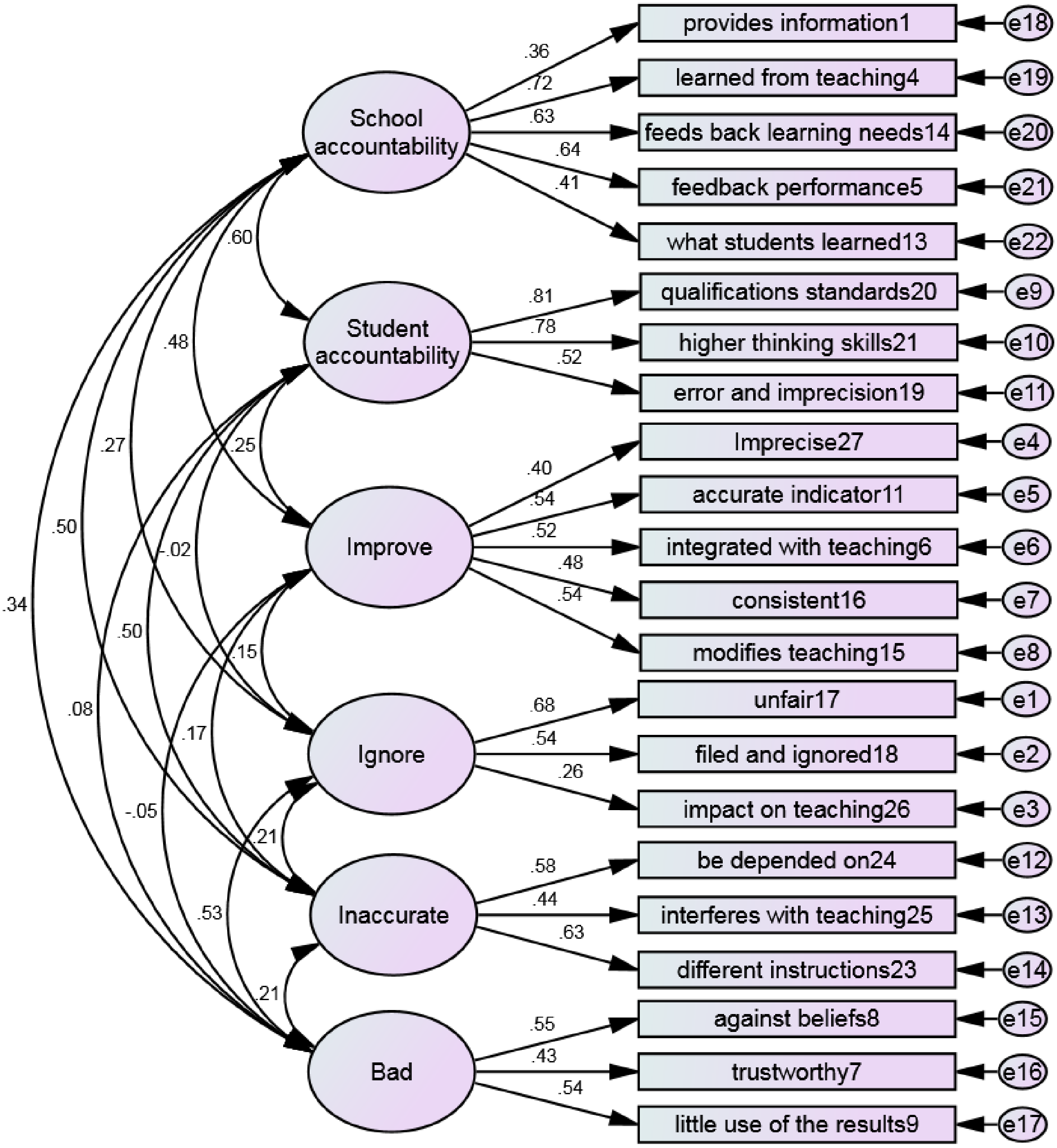
Teachers should take into account the error and imprecision in all assessment.
Assessment is an imprecise process.

Appendix B The Tunisian inadmissible model



Fit indices:χ^2: 1805.788, df: 312, $\chi^{2/df}$: 5.788, SRMR: .17, RMSEA: .094, CFI: .52, TLI: .46, all loadings were significant at p= .000

Appendix C A Tunisian model of conflicting indicators and factors **(**deselected items, 2,3, 10, 12 and 22)



Fit indices: χ^2: 525.278, *df*: 194, $\chi^{2/df}$: 2.708, SRMR: .10, RMSEA: .056, CFI: .84, TLI: .80, all loadings were significant at *p*= .000

Appendix D Deselected indicators from the data

1. *Assessment is an accurate indicator of a school's quality.*
2. *Assessment establishes what students have learned.*
3. *Assessment is integrated with teaching practice.*
4. *Assessment information modifies ongoing teaching of students.*
5. *Assessment results are trustworthy.*
6. *Assessment results are consistent.*
7. *Assessment results can be depended on.*
8. *Assessment is bad*
9. *Assessment interferes with teaching.*
10. *Assessment results should be treated cautiously given measurement error.*
11. *Teachers should take into account the error and imprecision in all assessment.*
12. *Assessment is an imprecise process.*