



Universiteit  
Leiden  
The Netherlands

## **Lost in translation : congruency of teacher and student perceptions of assessment as a predictor of intrinsic motivation in ethnodiverse classrooms**

Pat El, R.J.

### **Citation**

Pat El, R. J. (2012, October 9). *Lost in translation : congruency of teacher and student perceptions of assessment as a predictor of intrinsic motivation in ethnodiverse classrooms*. Retrieved from <https://hdl.handle.net/1887/19934>

Version: Not Applicable (or Unknown)

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/19934>

**Note:** To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/19934> holds various files of this Leiden University dissertation.

**Author:** Pat El, Ron Jonathan

**Title:** Lost in translation : congruency of teacher and student perceptions of assessment as a predictor of intrinsic motivation in ethnodiverse classrooms

**Issue Date:** 2012-10-09

## Validation of Assessment for Learning Questionnaires for Teachers and Students<sup>1</sup>

### Abstract

**Background:** Assessment can be a powerful force in promoting student learning. Still, few measures exist to gauge Assessment for Learning in the classroom. Literature on AfL suggests that it encompasses both a monitor to track student progress as well as a scaffold to show or help students recognize in what areas they need to improve.

**Aims:** Based on a review of recent attempts to measure the Assessment for Learning (AfL) we constructed Assessment for Learning Questionnaires for Teachers (TAFL-Q) and for Students (SAFL-Q) for evaluating perceptions regarding AfL practices in classrooms using matching items.

**Sample:** The total sample included 1,422 students (49% girls, 51% boys) and 237 teachers (43% females, 57% males) in lower vocational secondary education.

**Methods:** The 28-item questionnaires were examined by means of confirmatory factor analysis using EQS on one random half of the sample. The CFA was cross-validated on the second half. Measurement invariance tests were conducted to compare the student and teacher versions of the questionnaires.

**Results:** Confirmatory Factor Analysis revealed a stable second-order two-factor structure which was cross-validated: perceived monitoring, and perceived scaffolding subsumed under a common factor: AfL. Tests for measurement invariance showed that the parallel constructs were measured similarly for both students and teachers.

**Conclusion:** The TAFL-Q and SAFL-Q capture the construct AfL in two subscales: Monitoring and Scaffolding, and allows for comparisons between teacher and student perceptions. The instruments can be a useful tool for teachers and students alike to identify and scrutinize assessment practices in classroom.

---

<sup>1</sup>This chapter was published as: Pat-El, R.J., Tillema, H., Segers, M.S.R., & Vedder, P. (2011). Validation of assessment for learning questionnaires for teachers and students. *British Journal of Educational Psychology*.

## 2.1 Introduction

### Assessment for Learning

Current research into assessment as a tool to support student learning is increasingly focused on how this support is perceived. Studies focus either on the question of 'how' it is perceived or on 'how much' of it is perceived by students or instructors. There is evidence emerging that teachers and students differ in their perception of both the extent to which assessment is integrated in instruction as well as its content. However, quantitative instruments to properly compare student and teacher perceptions are still lacking.

It is generally accepted that assessment is a powerful tool to inform and support student learning. In this respect, the concept 'Assessment for Learning' (AfL) has been introduced to counterbalance the majority of the attention paid to 'Assessment of Learning', or performance assessment, in classrooms and schools (e.g., P. Black & Wiliam, 1998a; Crooks, 2001; Stiggins, 2005). In the past decades, several models have been proposed that integrate summative testing within the learning processes in formative ways; such as formative evaluations (Scriven, 2002) or informal feedback (Boud, 2000). Assessments have become a critical link between curriculum, teaching, and learning (National Council of National Council of Teachers of Mathematics, 1995). Central to AfL is the integration of assessment into instruction as an ongoing process, where assessment information is used by teachers to adjust their teaching and by students to adjust their learning processes (Popham, 2008). The definition of assessment itself has become very broad and an umbrella term for any information that can help teachers gain better insights into students' understanding. The notion of AfL integrates assessments of performance (process) and achievement (products) within the learning context (Segers, 2004). AfL is characteristic of an assessment culture, that builds on scaffolding, aligning instruction in support of learning, and providing students with greater autonomy in their learning processes (P. Black & Wiliam, 1998a). The AfL perspective revolves around repeated, varied and often informal assessments which are informative to the learning process (Wiggins, 1993), and take place in authentic contexts (Darling-Hammond, 2010).

Even though different scholars emphasize slightly different AfL principles (for example, Assessment Reform Group, 2002; P. Black & Wiliam, 1998a; James & Pedder, 2006), our focus here is on two factors, viz; a monitor to track student progress and a scaffold to show or help students recognize in what areas they need to improve (Stiggins, 2005). The importance of these two factors is stressed and clarified in the ARG (2002) guidelines as well as by the Council of Chief State School Officers (CCSSO) (Darling-Hammond, 2010) and were explored in several grounding reports and studies (e.g., OECD, 2005; Prosser & Trigwell, 1993).

Although there is a substantial body of quantitative research into effects of AfL on learning, the questionnaires used, especially those pertaining to the perception of practices of AfL as a tool to evaluate the promotion of learning through assessment, have rarely received detailed psychometric attention. This is the first challenge to be tackled in the current study.

## Perceptions of AfL

In order to understand how assessment influences student learning, research has addressed conceptual attitudes towards assessment practices (Brown, 2008) and perceived/observed practice (Gibbs & Simpson, 2003). Hirschfeld and Brown (2009) showed that in secondary, as well as in higher education, general conceptions of assessment purposes affect student attainment and well-being. Few studies have evidenced the influence of students' perceptions of applied assessment practices on their approach to learning (Entwistle, 1991; Struyven, Dochy, & Janssens, 2005). Teachers' perception of the assessment practice has received even less attention and the scant studies available show that teachers' and students' perceptions have very little overlap (Schultz, 2001). Such overlap, or alignment, may be difficult to achieve. Due to their different roles and expertise, students and teachers are likely to differ in the opportunities and focus of assessment. Könings (2007) suggests that teachers are more alert than their students to the perception of clarity of learning goals, motivating subject matter, and engaging learning environments for learning. Misalignment in the perceptions of assessment practices between teachers and students is problematic because it is likely to lead to misunderstanding and misinterpretation of the assessment information, its meaning, and purpose (Bartholomew et al., 2001). MacLellan (2001) and Loughran (2010) showed that in order to facilitate student learning, teachers need to make sure that at least teachers' and students' perception of assessment goals are in alignment. When the interactions between teachers and students are characterized by a misalignment, the difficulties involved in compensating or correcting this issue remain unclear. To facilitate early detection or to make sure that teachers' and students' perception of assessment situations and interactions are aligned it is indispensable to measure both teacher and student perceptions of AfL. Given these studies and the examples and arguments they present, we argue that it is important to distinguish between and measure both teacher and student perceptions of AfL practices in order to optimize the effects of assessment on student learning.

## Measuring Perceptions of Assessment for learning

Current instruments are, in our opinion, not adequate for the designed objectives. They either show a number of methodological shortcomings, such as lacking construct validity and low internal consistency of scales (e.g., Gibbs & Simpson, 2003), or do not allow the matching of student and teacher perceptions (e.g., MacLellan, 2001). For example, the Assessment Experience Questionnaire (Gibbs & Simpson, 2003) is a Likert-type questionnaire for use in higher education. It measures a variety of AfL constructs such as quantity, timing and quality of feedback in addition to students' study efforts as well as how students perceive the examination. However, this instrument showed low internal consistencies. Another example regards the questionnaire developed by James and Pedder (2006), which was constructed to measure primary and secondary school teachers' beliefs and appraisals of AfL practices using Likert-type responses. The instrument showed low factor loadings. Furthermore the discrimination between measured AfL constructs remains unclear. One can see that measuring congruency in perception between teachers and students has proven difficult.

From a measurement perspective, structural confirmations of the dimensionality of AfL constructs have not been tested. The key principles of AfL have not yet been captured instrumentally, nor were they quantitatively validated. Available instruments on AfL practices do not comprehensively capture how students and teachers perceive the interrelated AfL principles that are applied in their classrooms. This study aims to develop and structurally validate a questionnaire designed to measure student and the teacher perceptions of AfL and which further allows direct comparisons of teacher and student perception of their AfL practice.

## 2.2 Method

### Sample

The target population consisted of students and teachers in lower vocational education in the Netherlands. This educational track includes approximately 60% of the Dutch secondary school students (Central Bureau of Statistics, 2007). Students in the upper secondary educational system, those preparing for higher education, have a different pedagogical approach. This sample was chosen because review studies such as those from Black and Wiliam (1998a) indicate that the highest learning gains from using AfL are likely to be observed in the lower tracks. The individual students and teachers were the units of analysis. The questionnaire was administered in 80 schools to 1422 students (49% girls, 51% boys). From each school, three teachers were selected along with a random selection of six students from these classes. Teachers were selected from all disciplines, ranging from arts to sciences. The median student age was 14 years old (*range* = 11 - 18). We sampled 237 teachers (43% females, 57% males) who had a median age of 43 years old (*range* = 22 - 65). The classroom response rate was 98.75%. On average, the teachers had been active in the field of teaching for 15.3 years ( $SD = 11.11$ ) and active at their current school for 8.8 years ( $SD = 8.27$ ). Male teachers ( $Med = 45$ , *range* = 23 - 65) were slightly older than female teachers ( $Med = 37$ , *range* = 22 - 63,  $U = 5092.00$ ,  $p = .001$ ). The participating schools were spread across different regions in the Netherlands to avoid a bias towards urban areas.

### Instrument

Student Assessment for Learning Questionnaire (SAFL-Q) and the Teacher Assessment for Learning Questionnaire (TAFL-Q) consisted of 28 items divided into two scales: (1) perceived monitoring (16 items); (2) perceived scaffolding (12 items). The items were measured on a five-point Likert scale (1 = strongly disagree, and 5 = strongly agree) (see Table 2.2 and Appendix for items).

### Procedure

Schools were visited by research assistants who distributed the questionnaires in classrooms. Prior to administering the SAFL-Q and the TAFL-Q the purpose of the study was explained to teachers. The questionnaires started with a short instruction explaining the purpose of the study, presenting some examples on how to work with Likert response scales and directing teachers' and students'

attention to actual assessment practices in their current class. Completion took about 25 minutes.

### Data analysis

Because of the large number of respondents needed to validate the questionnaire, confirmatory analyses were only conducted on the SAFL-Q. The validation was conducted in three steps: (1) Confirmatory analysis to determine the robustness of the factor structure on one randomly determined half of the dataset (training); (2) cross-validation using multi-group invariance testing between the first half of the dataset and the unused second half of the dataset; and (3) tests of multi-group invariance to determine whether the teacher and student questionnaires measure parallel constructs similarly.

#### Confirmatory factor analysis (CFA).

To test the stability of the explored factor structure, CFA was performed on the first data set in structural equation modelling software: EQS version 6.1. To interpret a model's fit, the following indicators were used: Root Mean Squared Error of Approximation (RMSEA) and Standardized Root Mean Residual (SRMR) below 0.05 and Comparative Fit Index (CFI) scores above 0.95 indicate good fit (Browne & Cudeck, 1992) and RMSEA and SRMR below .08 and CFI scores above .90 indicate acceptable fit (Hu & Bentler, 1999). Because the  $\chi^2$  statistic becomes increasingly unreliable in sample sizes  $> 250$  (Byrne, 2006), a slightly less sample-dependent statistic was used:  $\chi^2$  divided by its degrees of freedom (df) (Bollen, 1989). CFI is penalized for model complexity, which means that in complex models, CFI might drop. A measure that does not penalize large or complex models such as CFI does, is the Gamma ( $\gamma$ ) (Fan & Sivo, 2007) which is calculated based on the number of manifest variables, df, and RMSEA, and should have values above .90 for acceptable fit and above .95 for good fit.

Four types of models were tested: (a) a uni-dimensional model; (b) an uncorrelated-factor model; (c) a correlated-factor model; and (d) a second order model with latent factors explained by a common factor.

#### Cross-validation.

To test the robustness of the explored factor structure, CFA performed on the training set was repeated on the test set and their invariance were tested using multi-group invariance testing.

#### Teacher-student invariance.

Increasing levels of equality constraints were applied to the tested models to assess structural invariance of the questionnaires between teachers and students. The tested types of invariance are (1) dimensional: are the same number of common factors present in each group?; (2) configural: are common factors associated with the same items across groups?; (3) metric (invariant first and second order factor loadings): do common factors have the same meanings across groups?; (4) strong factorial (invariant variable and factor intercepts):

are comparisons of group means meaningful?; (5) strict factorial (invariant error terms): are comparisons of group means and observed variables defensible (Gregorich, 2006)? Strict factorial invariance has been argued as too strict of a criterion (Byrne, 2006) and is generally not seen as a requirement for invariance between populations (Wu, Li, & Zumbo, 2007). A critical value of  $-0.01 \Delta CFI$  will be used to judge invariance (Cheung & Rensvold, 2002).

## 2.3 Results

### Confirmatory Factor Analysis

Confirmatory factor analysis was used (EQS version 6.1) to test the stability of the two-component structure of the 28 item SAFL-Q. No Heywood cases - negative estimations of error variances for an indicator of a latent variable - were found in any of the fitted models. A one-factor model was fitted (model A) which showed a weak fit ( $\chi^2/df = 3.61$ ; CFI = .85;  $\gamma = .92$ ; SRMR = .06; RMSEA = .06). This indicated that monitoring and scaffolding were correlated, but distinct factors. An uncorrelated two-factorial model with added factor-correlations was fitted (model B) which showed a weak fit ( $\chi^2/df = 3.44$ ; CFI = .86;  $\gamma = .92$ ; SRMR = .17; RMSEA = .06). A two-factor model with added factor-correlations (model C) showed a significantly improved and good fit ( $\Delta\chi^2(1) = 396.68$ ,  $p < .001$ ; CFI = .92;  $\gamma = .96$ ; SRMR = .04; RMSEA = .04). Finally, a second-order two-factorial model with a common AfL factor was fitted (model D) which showed an equally good fit ( $\Delta\chi^2(2) = 0.00$ ;  $p = 1.00$ ;  $\gamma = .96$ ; SRMR = .04; RMSEA = .04) which indicated that the two distinct factors monitoring and scaffolding were part of a single underlying construct. Table 2.1 summarizes the fit of the tested models. Table 2.2 shows the factor loadings of the common factor CFA.

Table 2.2: Second order two-factor CFA standardized factor loadings on 28-items and second-order factor loadings for Monitoring and Scaffolding with their common AfL-factor (loadings  $< .30$  are suppressed).

Item	CFA		Unique-ness
	monitor	scaffold	
1. My teacher encourages me to reflect on how I can improve my assignments	.56		.83
2. After examining my test results, my teacher discusses the answers I gave to the test with me	.39		.92
3. Whilst working on my assignments, my teacher asks me how I think I am doing	.53		.85
4. My teachers allow me to think about what I want to learn in school	.50		.86
5. My teacher gives me the opportunity to decide on my own learning objectives	.43		.90



Table 2.1: Fit indices of the SAFL-Q factor models.

<b>Model</b>	$\chi^2$	<b>df</b>	$\chi^2/\text{df}$	<b>SRMR</b>	<b>CFI</b>	$\gamma$	<b>RMSEA</b>	<b>90% CI</b>	<b>Model Comparison</b>	$\Delta\chi^2(\text{df})$
Model A: 1-factor	1269.33	351	3.61	.055	.845	.917	.060	.056, .064	—	—
Model B: 2-factor; uncorrelated	1204.67	350	3.44	.167	.856	.922	.058	.054, .062	—	—
Model C: 2-factors; correlated	807.98	349	2.32	.039	.923	.956	.043	.039, .046	C vs B	396.68**
Model D: 2-factors; with one second or- der common factor	807.98	347	2.33	.039	.922	.956	.043	.039, .047	D vs C	0.00

\*\* =  $p < .00$

6. My teachers inquire what went well and what went badly in my work	.62	.79
7. My teacher encourages me to reflect on my learning process and to think about how to improve next time	.67	.75
8. My teacher stresses my strengths concerning learning	.53	.85
9. My teacher identifies my weaknesses concerning learning	.37	.93
10. I am encouraged by my teacher to improve my learning process	.61	.80
11. My teacher gives me guidance to assist my learning	.60	.83
12. My teacher discusses assignments with me to help me understand the subject matter better	.56	.80
13. My teacher discusses with me the progress I make	.61	.78
14. After each assessment my teacher informs me how to improve the next time	.64	.77
15. My teacher discusses with me how to exploit my strengths to improve my assignment	.70	.71
16. My teacher and I consider ways to improve my weak points	.73	.69
17. When I do not understand a topic, my teacher tries to explain it in a different way	.58	.82
18. My teacher provides me with hints to help understand the subject matter	.65	.76
19. During class I have an opportunity to show what I have learned	.49	.87
20. My teacher asks questions in a way I understand	.67	.74
21. My teacher asks questions that help me gain understanding of the subject matter	.70	.72
22. My teacher allows for my contribution during the lesson	.56	.83
23. I have the opportunity to ask my classmates questions during the lesson	.43	.90
24. My teacher makes me aware of the areas I need to work on to improve my results	.44	.90
25. There is an opportunity to ask questions during the lesson	.50	.87
26. I am aware of the criteria by which my assignment will be evaluated	.54	.85
27. When I receive an assignment it is clear to me what I can learn from it	.62	.78
28. My assignments allow me to show what I am capable of	.47	.89
	AFL-common factor	
F1. Monitoring	.85	.53
F2. Scaffolding	.88	.47
Pearson $r$ Monitoring - Scaffolding = .63, $p < .001$		

---

Note: The translation is based on the Dutch original validated in this study

### Internal Consistency

Cronbach's  $\alpha$  of the two subscales in the teacher and the student questionnaire provided convincing support for the reliability of the subscales. Table 2.3 shows  $\alpha$ s and scale characteristics. The mean scores of teacher perceptions

were almost a point higher than student perception scores.

Table 2.3: Reliabilities and scale statistics

	Scale	Cronbach's $\alpha$	N	Mean	SD
TAFL-Q	Monitoring	.87	237	3.70	0.54
	Scaffolding	.77	237	4.35	0.38
SAFL-Q	Monitoring	.89	1244	3.06	0.74
	Scaffolding	.83	1244	3.82	0.64

### Cross-Validation

The training ( $N_1 = 726$ ) and the test ( $N_2 = 696$ ) sample subset were compared to test the robustness of the two-factor model. Multi-group comparisons in EQS, with varying levels of equality constraints, confirmed that the two-factor structure on the 28-items was equally robust in both random sets (Table 2.4), as the two randomly sampled halves were strictly factorially invariant. The baseline hypothesized model in the test sample ( $\chi^2/\text{df} = 2.33$  CFI = .92;  $\gamma = .96$ ; RMSEA = .04) was confirmed in the training sample ( $\chi^2/\text{df} = 2.28$ ; CFI = .92;  $\gamma = .96$ ; RMSEA = .04). The two samples proved strictly factorially invariant ( $\Delta CFI < .001$ ;  $\chi^2/\text{df} = 2.54$ ; CFI = .90;  $\gamma = .94$ ; RMSEA = .047).

### Teacher-student invariance

Structural equivalence of the two-factor structure between students ( $N = 1244$ ) and teachers ( $N = 237$ ) was tested in EQS by applying increasingly restrictive equality constraints. All equality constraints were correctly imposed. An overview of the steps and results is given in Table 2.5. Testing of the baseline hypothesized model for the full sample of students ( $N = 1422$ ) yielded a good fit to the data ( $\chi^2/\text{df} = 3.55$ ; CFI = .93;  $\gamma = .96$ ; RMSEA = .04). The Langrange Multipliers Test statistics, which help identify whether additional parameters should be added as well as Wald statistics, which help identify whether parameters should be removed, revealed that the fit for the student baseline model remained consistent with the hypothesized model. Testing for the validity of the TAFL-Q ( $N = 237$ ) yielded a weaker, yet still acceptable fit to the data ( $\chi^2/\text{df} = 2.41$ ; CFI = .74;  $\gamma = .86$ ; RMSEA = .08). Thus, with the exception of CFI, the fit measures indicate that the teacher baseline model is also consistent with the hypothesized model.

The common factor model was found to be strongly factorially invariant (Model 5). To prevent under-identification (more estimated parameters than observed) in the testing of the invariance of latent factor intercepts, the fixed intercepts were constrained to their observed intercepts values found in a separate model in which first- and second-order factor loadings were constrained equal across groups, however the observed variable intercepts were freely estimated. These values were .734 and .646 respectively. Testing revealed a good-fitting model ( $\chi^2/\text{df} = 4.23$ ; CFI = .93;  $\gamma = .93$ ; RMSEA = .05). Even though there

Table 2.4: Fit indices of multi-group comparisons between training and testing sample of the student dataset of the second order two-factor model

Model	$\chi^2$	df	$\chi^2/\text{df}$	CFI	$\gamma$	RMSEA	90% CI	Model Comparison	$\Delta\text{CFI}$
Model 1: Configurational	1847.62	694	2.66	.904	.946	.048	.046, .051	—	
Model 2: First-order factor loadings invariant	1871.95	720	2.60	.904	.942	.049	.045, .050	2 vs 1	<.001
Model 3: First- and second-order factor loadings invariant	1871.95	722	2.60	.904	.942	.049	.045, .050	3 vs 1	<.001
Model 4: Strict Factorial: Error variances invariant	1893.45	746	2.54	.904	.944	.047	.044, .049	4 vs 1	<.001

was a substantial change in CFI, it was in the direction of improved fit. Testing for the strict factorial invariance by constraining all estimated error terms resulted in a less than acceptable fit ( $\chi^2/df = 5.36$ ; CFI = .89;  $\gamma = .89$ ; RMSEA = .07). The strong reduction in fit between model 6 and model 1, combined with the  $\Delta$  CFI of .013, indicates that the TAFL-Q and the SAFL-Q are likely not strictly invariant.

## 2.4 Discussion

The aims of this study were to construct and validate a self-report questionnaire measuring teacher and student perceptions of ‘Assessment for Learning’ (AfL) practices in their classroom. Based on a review of the literature on AfL and drawing on the principles outlined by ARG and CCSSO, we developed two instruments to gauge AfL: the SAFL-Q, for students and the TAFL-Q for teachers. As expected, the results of our analyses show a good fit for a two-factor solution with 28 items that was robust for both the SAFL-Q and the TAFL-Q. The two-factors in the student and teacher questionnaires were labelled: ‘monitoring’, and ‘scaffolding’.

Strictly speaking, correlations showed that the two factors in both the SAFL-Q and the TAFL-Q are not independent. This is not surprising given that both practices in AfL, i.e., monitoring and scaffolding, are by their very nature linked. The empirically established two factors in this study clearly represent the constructs found in the literature. The factor ‘scaffolding’ has items that refer to clarification of learning goals and criteria and to classroom questioning, which are largely instruction-related processes. The factor ‘monitoring’ has items that deal with feedback and self-monitoring which share a common intent to optimize learning. However, the overlap is not complete. One reason might be that, although the important building blocks of AfL are relatively well-defined (Assessment Reform Group, 2002; Stiggins, 2005), they are not necessarily mutually exclusive or orthogonal. Nevertheless, the two factors cover many of the conceptually stated principles of AfL. We opted for an open approach of Assessment for Learning. It is not the assessment itself which determines whether assessment is summative or formative in nature (function), but rather how the assessment information is used (goal) (P. Black & Wiliam, 1998b). This means that feedback, even when summative in character, can become formative when its information is used to inform students how, and in what way, their performance falls short of a goal, and how to improve (Kluger & DeNisi, 1996). Although there is an ongoing debate as to which exact ingredients define assessment for learning (Wiliam, 2011), this study adds evidence to the suggestion (Stiggins, 2005; Pat-El, Segers, Tillema, & Vedder, 2008) that the reduction of dimensionality of the AfL construct can lead to more parsimonious overarching clusters, such as suggested by James and Pedder (2006, p. 110). This suggestion finds further basis on the works of Gipps (1994) in which AfL is called distinctive by its focus in particular on “learning as an explicit focus of classroom activity and students’ learning autonomy”, which we have summarized as activities to monitor student progress, and scaffolding activities to realize growth.

Another aim of our study was to establish questionnaires that would mea-

Table 2.5: Fit measures of multi-group comparisons between teachers and students

<b>Model</b>	$\chi^2$	<b>df</b>	$\chi^2/\text{df}$	<b>CFI</b>	$\gamma$	<b>RMSEA</b>	<b>90% CI</b>	<b>Model Comparison</b>	$\Delta\text{CFI}$
Model 1: Configural; No constraints	2095.06	694	3.02	.898	.944	.049	.047, .052	—	
Model 2: First-order factor loadings invariant	2236.28	720	3.11	.890	.940	.050	.048, .053	2 vs 1	.008
Model 3: First- and second-order factor loadings invariant	2236.28	722	3.10	.890	.939	.050	.048, .053	3 vs 1	.008
Model 4: First- and second-order factor loadings; observed variable intercepts invariant	3266.88	752	4.34	.896	.927	.054	.052, .056	4 vs 1	.002
Model 5: First- and second-order factor loadings; observed variable and latent factor intercepts invariant	3266.94	752	4.23	.925	.927	.054	.052, .056	5 vs 1	-.027
Model 6: First- and second-order factor loadings; observed variable and latent factor intercepts; error variances invariant	4189.99	778	5.39	.885	.892	.066	.064, .069	6 vs 1	.013

sure the perceptions of AfL practices in classrooms and could detect (possibly differing) perceptions of students and teachers. Previously available self-report research tools have not convincingly compared evaluative perceptions of actual practices of both populations –teacher and students- that are specifically related to promoting learning (MacLellan, 2001). The questionnaires in the current study used matching items and the measurement invariance between the two questionnaire versions indicate that the two-factor structure is not only robust, but can be similarly interpreted for both teachers and students alike. Our findings that corresponding latent factor intercepts are invariant across groups suggests that group differences in estimated factor means are unbiased and group differences in observed means are directly related to group differences in factor means and not contaminated by differential response bias (Gregorich, 2006). This means that, when used to compare perceptions of both student and teacher populations, differing perception scores can be attributed to different perceptions and not to different interpretation of the factors. The implication for researchers is that mean scores of teachers and students can directly be compared and difference scores can be used to measure the degree to which teachers' and their students' views about the perceived same classroom practices match without bias. Due to observations that students can perceive, for example, learning criteria to be more implicit and 'hidden' than their teachers believe they are, comparing student and teacher perceptions of AfL has become the focus of research trying to optimize the effects of assessment on student learning (Könings, 2007). If AfL is to integrate assessment with learning, it has to be a process of continual interaction between teachers and individual learners, in which feedback provision and its acceptance and utilization are key elements (P. Black & Wiliam, 2009). In this respect, it is of great importance that students' and teachers' perceptions on the nature and content of the assessment provided are congruent. Teachers may have to adapt word choice and complexity of the feedback-information to help students understand the feedback (P. H. Vedder, 1985). Conversely, it is important that students anticipate teacher instructions and feedback as personal expectations. In this manner, congruency in teacher and student assessment and learning related perceptions are important for the success of classroom interventions (Bartholomew et al., 2001; Loughran, 2010; Norman, 1986). Seeking discrepancies between student and teacher perceptions of AfL can help detect why some instructional activities to foster learning sometimes have unexpected or contradictory effects (Olkinuora & Salonen, 1992).

The instrument(s) in this study are more context-sensitive than overall measures of assessment conceptions as the questionnaires are designed to elicit responses on actual assessment practices closely related to the (promotion of) learning going on in classrooms. The results indicate much higher means in AfL perceptions for teachers over students. Reasons for this average misalignment could either be due to teachers being more knowledgeable about, or rather students being less aware of, the intent and purpose of their assessment. Students' perceptions in this sample are nested within classrooms and as such global comparisons of perception means will be biased. Comparisons that account for the multilevel nature of the relationships between specific teachers and students are needed.

Confirming the robustness of the outcomes in additional populations or different types of education would enhance the usefulness of the SAFL-Q and

the TAFL-Q. The population in this sample (lower vocational secondary education) can generalize to lower track students, who are being prepared for vocational education, however the sample cannot provide evidence for generalizability to students in higher tracks, who are prepared for higher education. Confirming invariance in the entirety of the Dutch secondary education would indicate its usefulness in broader contexts, such as countries where secondary education is less structurally diversified. At present we believe the instrument is a valid representation of AfL principles and able to gauge how assessment can contribute to learning. Based on our results we argue that the instruments can be useful tools for teachers and students alike in identifying and scrutinizing assessment practices in the classroom and eventually be supportive in establishing a common and shared assessment approach to scaffold learning in classrooms.



## 2.5 Appendix A. TAFL-Q items and subscales English.

### Monitoring

1. I encourage my students to reflect upon how they can improve their assignments
2. After a test I discuss the given answers with each student
3. While working on their assignments I ask my students how they think they are doing
4. I involve my students in thinking about how they want to learn at school
5. I give my students the opportunity to decide on their learning objectives
6. I ask my students to indicate what went well and what went badly concerning their assignments
7. I encourage students to reflect upon their learning processes and how to improve their learning
8. I inform my students' on their strong points concerning learning
9. I inform my students' on their weak points concerning learning
10. I encourage my students to improve on their learning processes
11. I give students guidance and assistance in their learning
12. I discuss assignments with my students to help them understand the content better
13. I discuss with my students the progress they have made
14. After an assessment I inform my students on how to improve their weak points
15. I discuss with my students how to utilize their strengths to improve on their assignment
16. Together with my students I consider ways on how to improve on their weak points

### Scaffolding

1. I adjust my instruction whenever I notice that my students do not understand a topic
2. I provide my students with guidance to help them gain understanding of the content taught
3. During my class, students are given the opportunity to show what they have learned
4. I ask questions in a way my students understand

5. By asking questions during class I help my students gain understanding of the content taught
6. I am open to student contribution in my class
7. I engage my students by discussing answers to assignments during class
8. I ensure that my students know what areas they need to work on in order to improve their results
9. I give my students opportunities to ask questions
10. My students know what the evaluation criteria for their work are
11. I ensure that my students know what they can learn from their assignments
12. I can recognize when my students reach their learning goals

Note. This English translation is based on the Dutch original validated in this study.

## 2.6 Appendix B. Items and subscales of the TAFL-Q (Dutch)

### Monitoring

1. Ik stimuleer mijn leerlingen om na te denken over hoe zij hun schoolwerk kunnen verbeteren.
2. Na het nakijken van een toets bespreek ik met iedere leerling zijn of haar antwoorden
3. Tijdens het maken van het schoolwerk vraag ik de leerlingen hoe ze vinden dat het tot nu toe gaat
4. Ik laat mijn leerlingen meedenken over de manier waarop ze willen leren op school.
5. Ik geef leerlingen de gelegenheid zelf hun leerpunten te bepalen
6. Ik vraag mijn leerlingen om aan te geven wat ze goed en minder goed hebben gedaan in hun schoolwerk.
7. Ik stimuleer leerlingen om terug te kijken op hun leerproces en om te bedenken wat ze een volgende keer beter anders kunnen doen.
8. Ik laat mijn leerlingen weten wat hun sterke punten zijn op het gebied van leren.
9. Ik laat mijn leerlingen weten wat hun zwakke punten zijn op het gebied van leren.
10. Ik moedig mijn leerlingen aan om hun leerproces te verbeteren.
11. Ik geef mijn leerlingen aanwijzingen die hen helpen bij het leren.
12. Ik bespreek het gemaakte werk met mijn leerlingen zodat zij de lesstof beter begrijpen.
13. Ik bespreek met mijn leerlingen hun vorderingen.
14. Na een toetsmoment vertel ik mijn leerlingen hoe zij hun zwakke prestaties kunnen verbeteren.
15. Ik bespreek met mijn leerlingen hoe zij hun sterke kanten kunnen gebruiken om hun werk te verbeteren.
16. Ik stel samen met mijn leerlingen een strategie vast om hun zwakke punten te verbeteren

### Scaffolding

1. Als ik merk dat leerlingen een onderdeel niet begrijpen pas ik mijn instructie aan.
2. Ik geef de leerlingen aanwijzingen om hen te helpen de lesstof te begrijpen.

3. Tijdens de les kunnen de leerlingen laten zien wat ze hebben geleerd
4. Ik stel de vraag op een begrijpelijke manier
5. Door de vragen te stellen tijdens de les help ik leerlingen om de lesstof begrijpen.
6. Ik sta open voor de inbreng van mijn leerlingen in de klas.
7. Ik ga met mijn leerlingen in discussie over de antwoorden.
8. Ik zorg ervoor dat de leerling weet aan welke punten hij of zij moet werken om zijn of haar resultaten te verbeteren.
9. Ik geef leerlingen de mogelijkheid om vragen te stellen.
10. De leerling weet aan welke eisen zijn of haar werk moet voldoen.
11. Ik zorg ervoor dat de leerlingen weten wat ze kunnen leren van een opdracht.
12. Ik kan zien of de leerling het leerdoel heeft behaald door zijn werk

## 2.7 Appendix C. Items and subscales of the SAFL-Q (Dutch)

### Monitoring

1. De leerkracht stimuleert mij om na te denken over hoe ik mijn schoolwerk kan verbeteren.
2. Nadat de leerkracht mijn toets heeft nagekeken bespreken we samen mijn antwoorden.
3. Tijdens het maken van mijn schoolwerk vraagt de leerkracht hoe ik vind dat het tot nu toe gaat.
4. De leerkracht laat mij meedenken over de manier waarop ik wil leren op school.
5. Ik krijg van de leerkracht de mogelijkheid om te bepalen wat mijn leerpunten zijn.
6. Mijn leerkracht vraagt mij wat ik goed en minder goed heb gedaan in mijn schoolwerk.
7. De leerkracht stimuleert mij om terug te kijken op mijn leerproces en om te bedenken wat ik een volgende keer anders kan doen.
8. De leerkracht geeft mijn sterke punten aan op het gebied van leren
9. De leerkracht geeft mijn zwakke punten aan op het gebied van leren.
10. Ik word aangemoedigd door mijn leerkracht om mijn leerproces te verbeteren.
11. Ik krijg aanwijzingen van de leerkracht die mij helpen bij het leren.
12. Mijn leerkracht bespreekt mijn gemaakte werk met mij zodat ik de lesstof beter begrijp.
13. Mijn leerkracht bespreekt met mij mijn vorderingen
14. Na een toetsmoment laat mijn leerkracht mij weten hoe ik het de volgende keer beter kan doen.
15. Mijn leerkracht bespreekt met mij hoe ik mijn sterke kanten kan gebruiken om mijn werk te verbeteren.
16. Samen met mijn leerkracht bedenk ik een manier om mijn zwakke punten te verbeteren.

### Scaffolding

1. Als ik de uitleg niet begrijp dan probeert de leerkracht het op een andere manier aan mij uit te leggen.
2. De leerkracht geeft mij aanwijzingen die mij helpen om de lesstof te begrijpen.

3. Tijdens de les kan ik laten zien wat ze hebben geleerd
4. De leerkracht stelt de vraag op een begrijpelijke manier.
5. De vragen van de leerkracht helpen mij de lesstof te begrijpen
6. De leerkracht staat open voor mijn inbreng in de klas.
7. Ik heb de mogelijkheid om vragen te stellen aan medeleerlingen over de les.
8. Ik weet aan welke punten ik moet werken om mijn resultaten te verbeteren.
9. Er is mogelijkheid om vragen te stellen.
10. Ik weet aan welke eisen mijn werk moet voldoen.
11. Als ik een opdracht krijg is het duidelijk wat ik hiervan kan leren.
12. Met mijn werk laat ik zien wat ik kan.