**Swedish Teacher Conceptions of Assessment: A focus on improving outcomes**

Gavin T L Brown1,2, Catarina Andersson3,4, Mikael Winberg3,5, Björn Palmberg3,4, & Torulf Palm3,4

1 Department of Applied Educational Science, Umeå University, Sweden

2 Faculty of Education & Social Work, The University of Auckland, New Zealand

3 Department of Science and Mathematics Education, Umeå University, Sweden

4 Umeå Mathematics Education Research Centre (UMERC), Umeå University, Sweden

5 Umeå Science Education Research (UmSER), Umeå University, Sweden

Correspondence concerning this paper to Prof. Gavin T L Brown, Educational Measurement, Dept. of Applied Educational Science, Umeå University, Sweden or by email to [gavin.brown@umu.se](mailto:gavin.brown@umu.se)

Author Orcid numbers

Brown: 0000-0002-8352-2351

Andersson: 0000-0002-8738-8639

Winberg: [0000-0002-1535-873X](https://orcid.org/0000-0002-1535-873X)

Palmberg: 0000-0001-7089-0246

Palm: 0000-0003-0895-8232

**Keywords:** Sweden; Conceptions of Assessment; Teachers; Invariance; Confirmatory Factor Analysis

Manuscript submitted for publication. This is the author preprint version.

**Abstract**

Understanding teachers’ conceptions of assessment is a key objective in supporting assessment practices that lead to improved learning outcomes. Thus, inventories capable of identifying teachers assessment conceptions are important. The Teachers Conceptions of Assessment (TCoA) inventory was an early and influential measure of teacher assessment conceptions, but replication studies have shown that the model may be affected by policy and practice context. In the present study, a Swedish adaptation of the TCoA was administered twice, 18 months apart. A sample of 257 teachers were matched across the two time-points and their self-reported scores were analysed with confirmatory factor analysis and invariance testing. With good correspondence to the data, five of the nine factors in the TCoA were completely replicated and one factor was partially replicated. The model had sufficient similarity between time points to permit mean score comparisons, which were largely equivalent between times. The study indicates that the Swedish Teacher Conceptions of Assessment adaptation can be used reliably in Swedish primary and lower secondary schools as measure of teacher conceptions of the uses of assessment.

**Introduction**

Given the importance of teacher beliefs as predictors of practice (Fives & Buehl, 2012), knowing how teachers conceive of assessment seems to be essential in understanding their role in implementing assessment practices that enhance learning outcomes. Therefore, inventories capable of identifying teachers' assessment conceptions are needed. If teachers conceive of assessment purely in terms of summative evaluation of schools or teachers, it is unlikely that they will implement assessment practices that lead to improved learning outcomes. The Teachers Conceptions of Assessment (TCoA; Brown, 2003) inventory, developed in New Zealand, was an early and influential measure for exploring teacher thinking about the purposes of assessment. It has been evaluated in multiple educational jurisdictions and languages leading to the understanding that only when the policy and practice context is similar to that of New Zealand does the inventory elicit similar responses. To date no study in Sweden, a low-stakes assessment context, has been conducted. This paper contributes to that gap by examining in which ways the New Zealand model replicate in the Swedish context. We do so with a repeated measure invariance study of Swedish TCoA responses. Their responses produce partial replication of previously reported results from other low-stakes contexts, supporting the importance of ecological rationality in understanding teacher’s conceptions of assessment.

**Formative assessment in classroom contexts**

Formative assessment that is aligned to curricular objectives and goals requires teachers to collect information about student progress, difficulties, and successes to plan instruction, design activities, ask questions, and provide feedback to learners (Bloom, Madaus, & Hastings, 1981). Performance data can be interactions in a classroom (e.g., question and answer) but also includes more formal diagnostic testing (Brown & Hattie, 2012) or analysis of errors made in classroom or home practice (Bejar, 1984). From these kinds of performance data teachers can formatively make appropriate adjustments to their classroom instruction and to student learning activities (Lai & Schildkamp, 2016).

Naturally, this approach to assessment requires resources (i.e., assessments that diagnose needs and time to plan responses), policies that prioritise using assessment formatively rather than solely for administrative or summative purposes, and teacher commitment to using assessment formatively for improvement. Further, considering the importance of the teacher’s active role in using performance data in this way, it is likely that teacher beliefs about assessment and feedback matter to the efficacy of these processes.

Ajzen’s (1991) Theory of Planned Behaviour (TPB) identifies the importance of teachers’ beliefs and attitudes toward any phenomenon (in this case, assessment) as essential to understanding their intentions and actions. TPB also points to the importance of social norms and perceptions of behavioural control as predictors of intentions, behaviours, and outcomes. This aligns well with Fives and Buehl’s (2012) model in which teacher beliefs act as filters, frames, and guides to cognitive resources that impact their actions. Of course, teacher beliefs about the proper role of assessment are not universal; they are context bound by the policy and practice framework in which they are employed (Bonner, 2016; Brown & Harris, 2009; Fulmer, Lee, & Tan, 2015).

**Teacher Conceptions of Assessment**

Research into teacher beliefs about assessment indicates that teachers have multiple conceptions of assessment with varying degrees of intensity (Brown, 2008). The focus of a belief (e.g., the type of assessment [formal vs. informal] and the object of evaluation [teacher/school vs. student]) matters to how strongly a conception of assessment is endorsed or rejected, and even to whether that conception is present in teachers’ conceptions of assessment. For example, in both New Zealand and Queensland, Australia, primary school teachers tended to not agree as much as secondary teachers did with the purpose of holding students accountable through assessment (Brown, 2011; Brown, Lake, & Matters, 2011). Arguably, this reflects the greater use of assessment among adolescents to prepare them for entry to further education or adulthood. Thus, because of the multiple purposes and uses of assessment, teachers have multiple and complex attitudes or conceptions in response to those uses within a jurisdiction.

It appears teachers are aware of formative and summative uses of assessment and their attitudes toward such uses vary depending on whether the student or the school is the focus of attention (Harris & Brown, 2009). The use of student assessment to evaluate schools and teachers especially creates tensions and negative responses from teachers (Bonner, 2016; Nichols & Harris, 2016). Teachers in high-stakes accountability systems, where public examination of student achievement have considerable acceptance, seem to view assessment for making students accountable a positive aspect for improving both student learning and character (Brown et al., 2009; Brown et al., 2011). The relative strength of these varying conceptions appears to be ecologically rational in that teachers in general endorse the policies and purposes that apply to their level of employment.

Hence, we should expect that Swedish teachers’ responses to the TCoA to reflect broad cultural values and specific educational policies about the nature, purpose, and practice of assessment.

**The Swedish Context**

According to the collaborative European Values Study and World Values Survey spanning 2005 to 2022 (EVS/WVS, 2022), Sweden exhibits a distinct inclination towards secular-rational and individualistic values, emphasizing equality, individual freedom, and well-being. This sentiment is reinforced by the six-dimensional Index of National Culture (INC) developed by Hofstede et al. (2010). Within the educational context, these values manifest in the Swedish system by granting teachers significant autonomy to interpret and implement the objectives outlined in the national curriculum. This extends to their ability to select appropriate teaching methods that align with the overarching goals of fostering student achievement (Helgøy & Homme, 2007).

The egalitarian ethos prevalent in Sweden extends to the students. Instances of discrimination among students, such as the creation of specialized classes or educational tracks for gifted or underperforming students, are infrequent (Hofverberg & Winberg, 2020; Hofstede et al., 2010). Practices like public evaluation and ranking of students based on academic achievement are generally considered unacceptable within the Swedish educational culture.

The Swedish curriculum adopts a goal-oriented approach, incorporating national standards for student achievement in Years 3, 6, and 9 (corresponding to ages 9, 12, and 15). While achievement grades are assigned in Years 6-9, the most significant implications lie in Year 9, influencing admission to upper secondary school. Grading is criterion-referenced, allowing multiple students to receive the same grade as long as they meet the established standards. Legally, the responsibility for providing adequate resources for education and conducting systematic evaluations lies with the municipality where the school is situated, as stipulated by The Education Act, (SFS, 2010:800, ch. 4 § 3). To support schools and teachers, the National Agency for Education (NAEd) offers national screening materials, assessment support, and standardized national tests. These tools serve diverse purposes, from informing decisions about support and guiding adaptations in teaching to providing aggregated estimates of student achievement at the school or system level for grading consistency and trend analysis (Skolverket, 2020).

National standardized tests (NSTs) are mandatory in Years 3, 6, and 9 and play a pivotal role in grading students in Years 6-9, although teachers are obligated to also consider other available information about students’ knowledge and skills in their evaluation process (SFS, 2010:800, ch. 10, § 20a). Consequently, the NAEd advises teachers to employ diverse assessment situations for both formative and summative purposes (Skolverket, 2022), granting teachers across all school levels the autonomy to design and implement their own classroom assessments.

Despite limited research on Swedish teachers' views on assessment and feedback, insights from Helgøy and Homme's (2007) study, involving approximately 70 teachers and principals across seven schools in four major cities in Sweden and Norway, indicate that Swedish teachers, more so than their Norwegian counterparts, perceive national standardized tests as valuable tools for grading and enhancing teaching. In contrast to Norwegian teachers, Swedish educators do not view these tests as constraining their autonomy in interpreting national goals or organizing their teaching methods to align with those goals.

Thus, we might expect a tendency among Swedish teachers to embrace a commitment to improvement-oriented conceptions on assessment, while also harbouring positive attitudes towards formal, standardized testing, acknowledging its possible uses for formative purposes. Furthermore, it is conceivable that teachers in closer proximity to the Year 9 decision point for entry into senior high school may demonstrate somewhat different belief priorities compared to their counterparts at the early stages of schooling.

**Methods**

This study used self-administered, self-reported survey inventories administered with a forced-choice ordinal agreement response scale at two time-points separated in time by eighteen months to check model replicability over time (reliability). Analysis was done within the multiple indicators, multiple causes (MIMIC; Jöreskog & Goldberger, 1975) framework in which each survey item response is explained by a latent factor and a residual capturing the universe of unexplained variance. Further, each latent construct is manifested by multiple indicators. This study uses confirmatory factor analysis and invariance testing to establish the structure of responses and their stability over time. Furthermore, the study examines whether the strength of conceptions differed depending on teacher sex, possession of a bachelor’s degree, length of teaching experience, or teaching stage.

**Participants**

A total of 461 teachers working between years 1 and 9 in a northern city in Sweden responded to the first wave of the survey. This is a 62% response rate from the municipality. A similar size group responded in the second wave of the survey held approximately 18 months later. After restricting the sample to those with complete data in both waves or those with <10% missing responses, 257 participants could be matched (Table 1).

<Insert Table 1 about here>

Consistent with school employment statistics, nearly ⅘of the sample were women and a similar proportion had a bachelor’s degree. Nearly ⅔ of the sample had more than 10 years teaching experience. Almost a third taught Years 1-3, a fifth in Years 4-6, and a quarter taught Years 7-9. Teachers indicated their specialist teaching subjects, with the option to indicate more than one subject. However, it is not possible to separate teachers into separate teaching subjects since in Y1-6 teachers are generalists teaching multiple subject areas. This means that, without studying exclusively teachers working above Year 6, it is not possible to identify possible subject or discipline related differences.

**Teachers Conceptions of Assessment (TCoA)**

The Teachers Conceptions of Assessment (TCoA; Brown, 2003) inventory probes three dominant perspectives of assessment: (1) improvement of both teachers’ instruction and students’ learning on the grounds that assessment provides dependable diagnostic information; (2) demonstrating through assessment results that schools and students have fulfilled expectations; and (3) assessment is irrelevant or ignored because it is inaccurate or has bad effects (Brown, 2004). The New Zealand model comprises four correlated factors (i.e., Improvement, Irrelevance, School Accountability, and Student Accountability) and the first two factors had multiple subordinate factors (i.e., assessment is reliable, assessment is descriptive, assessment helps teachers, and assessment helps students within the improvement factor; assessment is bad, assessment has error, and assessment is ignored within the irrelevance factor). Each of the nine factors had three items. The TCoA was developed in the low-stakes assessment context of New Zealand with primary school teachers and achieved replication with primary teachers in Queensland, Australia, (Brown, 2006) and secondary teachers in New Zealand (Brown, 2011). Previous studies show that teachers endorse the improvement function, reject the irrelevance and assessment for school accountability conceptions, and have ambivalent views toward student accountability, with only secondary school teachers giving a somewhat positive endorsement of assessment for student accountability (Brown, 2017). Nonetheless, the societal norms around high- vs. low-stakes uses of educational assessment impact teacher conceptions of assessment. Previous research comparing the TCoA survey results in Cyprus, Queensland, India, Hong Kong, China, Egypt, Spain, and Ecuador found that the New Zealand model was replicated only among primary teachers in Queensland (Brown, Gebril, & Michaelides, 2019), a context that shared the English language, an assessment for learning pedagogy, and low-stakes consequences for assessments. Hence, although the present survey uses primarily primary school teachers in a low-stakes assessment for learning context, it is still likely that linguistic and policy factors may necessitate modifications in the statistical model representing teachers’ conceptions of assessment to gain close correspondence with the data.

**Survey Design**

The instruments were translated into Swedish by a translator, prioritising functional equivalence rather than literal equivalence. The translation was evaluated by three independent bilingual researchers who reported that the Swedish version had strong functional equivalence to the English.

A member checking (Tong et al., 2007) was conducted in two interview groups to find out how the teachers perceived certain items and understand the reasons why they answered the way they did. Ten teachers were recruited on a volunteer and convenience basis to participate. At separate occasions two group interviews (n = 4 and 6, respectively) were led by two of the authors. The interviews were focused on two items that indicated ambiguous teacher interpretation during analysis. In the interviews, the teachers were first given time to think individually about each issue and then shared their thinking in a joint discussion. The teachers had completed the survey, but their responses had been anonymous and so they were commenting on aggregate data results to which they had contributed. In these 1-hour discussions, we explored the respondents’ interpretations and views of the item content. The authors took field notes and conversations were audio-recorded and transcribed. The transcripts and field notes were analysed for themes by authors 2 and 5.

Items were presented in jumbled order. Participants responded using a positively packed, 6-point agreement scale. This type of scale has two negative options (Strongly Disagree and Moderately Disagree, scored 1 and 2 respectively) and four positive options (Slightly Agree, Moderately Agree, Mostly Agree and Strongly Agree, scored 3 to 6, respectively). This approach gives greater ability to discriminate the degree of positivity participants hold for positively valued statements and is appropriate when participants are likely to endorse statements (Brown & Shulruf, 2023). Hence, in circumstances when participants are expected to respond positively to a stimulus (e.g., teachers responding to a policy expectation), giving them more choices in the positive part of the response continuum produces good results.

**Analysis**

After deleting participants with >10% missing responses, and thereafter testing whether missing data were Missing Completely at Random (MCAR), the remaining missing values were imputed with the expectation maximisation algorithm (Dempster, Laird, & Rubin, 1977). This process creates a data set with no missing values for variance-covariance structural analysis. A statistically not significant result for Little’s (1988) Missing Completely at Random (MCAR) test or for the normed chi-square alternative (Wheaton et al., 1977) shows that the distribution of missing responses is completely random. At time 1, missing values ranged from none to 1.9%, with Little's MCAR test not statistically significant at the normed chi-square level (χ2=512.67, *df* = 455, *p* = .03; χ2/*df =* 1.13, *p* = .29). At time 2, missing values ranged from none to 2.3%, with Little's MCAR test not statistically significant at the normed chi-square level (χ2=728.86, *df* = 566, *p* < .001; χ2/*df =* 1.29, *p* = .26). Hence, the missing value imputation was accepted as a valid representation of teacher intended responses.

Just over half of the variables (30/54; 55%) violated normality for skewness (z > 3.29; Kim, 2013), while just six variables exceeded normality for kurtosis. All variables were corrected using the Box-Cox normalisation method implemented in Courtney and Chang’s (2018) normal R shiny app[[1]](#footnote-1). With this adjustment, normal theory covariance structure analysis was permissible (Bentler & Dudgeon, 1996).

In the event a CFA model is admissible, but poorly fitting, modification indices can be consulted to identify items that violate simple structure (Revelle & Rocklin, 1979) or independence of residual assumptions (Barker & Shaw, 2015), while maximising recovery of the original model. Items with weak loadings (i.e., <.30; Bandalos & Finney, 2010) or with strong violation of model assumptions (i.e., loading on non-intended factors or item residuals) are good candidates for deletion.

Fit of CFA models is established by inspection of multiple fit indices (Hu & Bentler, 1999). Because the chi-square measure of discrepancy between a model and its underlying data is sensitive to sample size and model complexity, statistically non-significant values for the normed chi-square (i.e., χ2/*df*) provide evidence for correspondence between a model and its data (Wheaton et al., 1977). Further evidence for non-rejection of a model arises when the comparative fit index (CFI) is >.90 and the root mean square error of approximation (RMSEA) is <.08. However, both the CFI and RMSEA indices are sensitive to models with more than three factors, with the CFI tending to reject and the RMSEA tending to not reject under those conditions (Fan & Sivo, 2007). While the standardized root mean residual (SRMR) also varies with model conditions, that index can be relied upon even in large models with non-normal data, although the sample SRMR will produce values greater than the population SRMR, suggesting that the model fits the data more poorly than it actually does (Maydeu-Olivares, Shi, & Rosseel, 2018). Thus, greater reliance is put on the gamma hat >.90 and the standardized root mean residual (SRMR) <.08 because these are robust against sample size, model complexity, and model misspecification.

Scale reliability was estimated using the Coefficient H maximal reliability index, which is based on an optimally weighted composite using the standardised factor loadings (Hancock & Mueller, 2001). H > .80 provides strong evidence for the robustness of a scale.

In order to evaluate the impact of time on survey responses, invariance testing between two time points was conducted. Nested invariance testing establishes if the item regression weights (i.e., metric equivalence) and intercept values (i.e., scalar equivalence) are equivalent across time. This method tests equivalence sequentially and stops if the additional constraint causes a change in the CFI index of more than .01 (Brown et al., 2017) or if a statistically significant result for the difference in chi-square, given the difference in degree of freedom, occurs. The equivalence of measurement weights is sufficient to meet requirements for creating and evaluating repeated scale scores over time (McArdle, 2007). However, if strict invariance is observed (i.e., equivalent weights at measurement and structural levels), we can conclude that the passing of two years did not change participant responses.

If at least metric invariance over time is achieved, factor means can be calculated by averaging the raw score for each item contributing to the factor. This unrefined method preserves the participant’s response scale metric producing easy interpretation of scores and is appropriate when simple structure (i.e., items belong to only one factor) is present (DiStefano, Zhu, & Mîndrilă, 2009).

CFA and invariance analyses were conducted in IBM SPSS Amos v29 (Amos Development Corp., 2022). Because six-point ordinal scales function similarly to continuous variables (Finney & DiStefano, 2006), maximum likelihood estimation was used.

**Results**

**Model development**

The original four-factor, hierarchical model described above (Brown, 2003) was not admissible because the correlation matrix among the four main factors was not positive definite. The Student Accountability factor had low prediction values loadings (i.e., <.30) for two items. Consequently, the whole factor was removed from the model. This produced an admissible but somewhat poorly fitting model.

Detailed inspection of the Irrelevance factor suggested several items violated simple structure, requiring further trimming. Item ir3 (taking account of error and imprecision in assessment) within the irrelevance factor had weak loading (<.30) and so was removed. Items ir2, ir1, ir3, ir5, and ir6 all had strong modification indices to items in other factors and were removed. This left Irrelevance with four items (ig1, ig2, ig3, and ir4).

In contrast, the hierarchical Improvement factor was fully replicated as four sub-factors each with three items. The School Accountability factor also was retained with its three items on assessment indicating school quality. Table 2 provides factors, items in both Swedish and English, and loadings for both time-points.

<Insert Table 2 about here>

The model (Figure 1) consisted of 19 items in a correlated three-factor structure (i.e., Irrelevance, School Accountability, and Improvement), with four subordinate factors under Improvement (Figure 1, Table 1). This model had qualitatively acceptable to good fit indices (χ2= 274.01, *df*= 145, χ2/*df*= 1.89, *p*= .17; CFI = .90; gamma hat = .95; RMSEA = .059, 90%CI = .048-.070; SRMR = .058) and so was not rejected. While the relatively low individual scale Coefficient H values raise doubts about the robustness of the measures, it is worth noting that the overall fit indices from confirmatory factor analysis support acceptance of this multi-dimensional self-report measure.

<Insert Figure 1 about here>

**Invariance**

The sequential testing of constraints, based on the change in CFI (i.e., ΔCFI range: .000 - .004), indicated full equivalence. However, the difference in chi-square by difference in degrees of freedom testing indicated that equivalence stopped after constraining measurement weights to be equivalent (Δχ2 = 8.395, Δ*df* = 3, *p* = .039). The fit of this two-group, constrained model was good, especially for the more robust normed chi-square, gamma hat, and SRMR indices (χ2= 792.26, *df*= 310, χ2/*df*= 2.56, *p*= .11; CFI = .85; gamma hat = .95; RMSEA = .055, 90%CI = .050-.060; SRMR = .058).

Under the assumption of full equivalence, the scale inter-correlations varied by chance across time waves. At both times, Improvement was positively correlated with School Quality and inversely with Irrelevance, suggesting that the teachers perceived that assessment was relevant for improvement and showing that good quality schools cause improvement to learning with assessment.

<insert Table 3 about here>

**Scale Scores**

In general, mean scores (Table 2) were close to or slightly above moderate agreement for assessment helps teachers and students improve and assessment provides descriptive, diagnostic information. In contrast, both assessment provides reliable, valid information and assessment identifies school quality had means between slightly and moderately agree. Unsurprising, the means for assessment is irrelevant were just below slightly agree.

Univariate ANOVA of mean scores by time showed a statistically significant decrease for assessment helps students improve (*F*(1,517) = 25.25, *p* < .001, η2partial = .047) and a statistically significant increase for assessment provides valid, reliable information (*F*(1,517) = 13.72, *p* < .001, η2partial = .026). It is worth noting that both of these score differences qualify as small effects (Cohen, 1988). All other factors were statistically not significant in means across the two time points. Hence, score differences over time were tiny or nil. Note that all means fell between slightly agree and moderately agree, with somewhat large standard deviations, suggesting averages were modest with room for substantial improvement.

Multivariate analysis of variance, using main effects and all two-way interactions, showed no statistically significant difference in scale means at either time point by teacher sex, possession of a bachelor’s degree, length of teaching experience, or teaching stage. Post-hoc Tukey HSD tests indicated several factor differences according to teaching stage, with teachers in Years 7-9 standing out (Table 4). At time 1 the teachers of Y7-9 students agreed assessment was more validly reliable than the teachers of Y4-6 (*p* = .046, 95% C.I. = .01, .82). At time 2, there were three distinctions between the Y7-9 teachers and either one or both other groups:

* Y7-9 teachers were higher for assessment helps students improve compared to both groups of teachers (Y4-6: *p* = .03, 95% C.I. = .04, .92; Y1-3: *p* = .01, 95% C.I. = .11, .86),
* Y7-9 teachers were higher for assessment provide diagnostic descriptions of performance compared to the early primary years teachers (Y1-3: *p* = .01, 95% C.I. = .10, .84), and
* Y7-9 teachers were lower for assessment is irrelevant compared to the early primary years teachers (Y1-3: *p* = .002, 95% C.I. = -1.04, -.19).

It would appear then that the subject specialist teachers working in lower secondary schools (Y7-9) had greater confidence in formative assessment conceptions than their generalist counterparts, especially those teaching in early primary school years.

<Insert Table 4 about here>

**Discussion**

A Swedish adaptation of the New Zealand Teacher Conceptions of Assessment inventory (version IIIA) was administered twice, approximately 18 months apart, in one northern Sweden municipality. A sample of 257 teachers were matched across the two time-points and their self-reported scores were analysed with confirmatory factor analysis and invariance testing. With good correspondence to the data, five of the nine factors in the TCoA-IIIA were completely replicated and the Irrelevance factor was partially replicated. The model had sufficient similarity between time points to permit mean score comparisons, which were largely equivalent between times. Small differences were identified for just a few scales for the Y7-9 teachers.

The student accountability factor identified in previous studies was completely not present in this study. Unlike the findings of a New Zealand interview study (Harris & Brown, 2009) in which teachers were very much opposed to the idea of holding students accountable, these Swedish teachers could not even systematically reject the conception. Instead, it is as if the conception is simply non-existent. Within the Swedish context, the items for this construct seem to violate Swedish cultural norms and educational policies. For example, item sa1 (Assessment places students into categories) can be interpreted as measuring and sorting students themselves, which is not socially acceptable in Sweden’s egalitarian culture. Both items sa2 (Assessment is assigning a grade or level to student work) and sa3 (Assessment determines if students meet qualifications standards) target students’ competencies in relation to qualification standards. However, the national steering documents require that assessment of students’ competencies in relation to qualification standards should be done both formatively and summatively, a view also expressed in the member checking interviews. Further, item sa2 specifically concerns the assignment of grades, which are only given in Years 6 to 9. Given that the current sample had a majority of teachers working in earlier years, this may explain why the item did not work.

The trimming of items from the irrelevance factor removed three items having to do with the need for teachers to take into account the degree of error in any assessment or be cautious with potential misinformation from or within assessment (i.e., ir2 Assessment results should be treated cautiously because of measurement error; ir3 Teachers should take into account the error and imprecision in all assessment; and ir6 Assessment is an imprecise process). Information about the degree of uncertainty or imprecision in resulting scores is a characteristic most associated with standardised tests that can be used for formative purposes. Perhaps, when teachers use diverse (unstandardized) assessment methods, and also base their assessments on their own interaction with a class or an individual, judgements about error might not only be difficult to make but are also unlikely to happen. Two other items that were deleted had to do with assessment having a negative or inappropriate influence upon teaching practices (i.e., ir1 Assessment forces teachers to teach in a way against their beliefs and ir5 Assessment has little impact on teaching). Given the inverse relationship between the Improvement factor and the Irrelevance factor it is likely that, among these Swedish classroom teachers, assessment is embedded within teaching and does express the teacher’s curricular goals and pedagogical choices. Further, given the control Swedish teachers have over assessment practices, it is unlikely that they would consider assessment to have little impact on their teaching. The very essence of formative assessment is that assessment derives from teaching and informs further instruction. Hence, these two statements could be seen as redundant and contradictory.

Swedish teachers generally agreed that assessment helps teachers and students improve and provides diagnostic descriptions of student performance. However, there is much room for increased endorsement of these formative conceptions. That goal might be addressed by providing assessment tools and protocols that build on those objectives, an approach taken in New Zealand with the deployment of the Assessment Tools for Teaching and Learning computer adaptive system (Brown, 2013, 2019; Brown & Hattie, 2012). Since the study indicates that the Swedish adaptation of the TCoA can be used reliably in Swedish primary and lower secondary schools as measure of teacher conceptions of the formative uses and purposes of assessment, it may be used in studies examining the need for, and effects of, professional development initiatives to support the development of teachers’ formative assessment conceptions.

Some limitations of the study can be identified. The measurements were performed in a single municipality in northern Sweden, which may reduce generalizability of the results. However, we posit that the findings here are likely representative of a broader cohort of Swedish teachers. While the school administration is local, national policy as manifested in, for example, national curriculum documents, national tests, guides for grading, and other national support material are the same nationwide in primary and lower secondary schools. However, assuring generalizability would require a national survey. The relatively modest coefficient H values (i.e., all H < 0.80) indicate low stability. Further testing of the factor measurement models in an additional sample of Swedish teachers is therefore warranted. Since the data constitutes responses to a questionnaire, we made informed interpretations, corroborated by a small-scale member checking exercise, of the meaning of factors and path values. Subsequent more comprehensive qualitative studies exploring teachers’ interpretations of the questionnaire items by providing them with their own scores could provide further confidence in our explanations.

This study indicates that the Swedish Teacher Conceptions of Assessment adaptation can be used reliably in Swedish primary schools as a measure of teacher attitudes toward the uses of assessment in that context. The near replication of the New Zealand model may be understood in the light of the similarities of the Swedish and New Zealand contexts. Both have low-stakes assessment systems and emphasis on formative uses of assessment. However, the differences between the New Zealand and Swedish models found in this study underscores the need for validation (and, possibly, adjustments) of the inventory even when educational systems share many policy features.

**References**

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*, 179-211. https://doi.org/10.1016/0749-5978(91)90020-T

Amos Development Corp. (2022). *IBM SPSS Amos*.In (Version 29.0.0) IBM. <https://amosdevelopment.com>

Bandalos, D. L., & Finney, S. J. (2010). Factor analysis: Exploratory and confirmatory. In G. R. Hancock & R. O. Mueller (Eds.), *The Reviewer's Guide to Quantitative Methods in the Social Sciences* (pp. 93-114). Routledge.

Barker, L. E., & Shaw, K. M. (2015). Best (but oft-forgotten) practices: Checking assumptions concerning regression residuals. *The American Journal of Clinical Nutrition*, *102*(3), 533-539. <https://doi.org/10.3945/ajcn.115.113498>

Bejar, I. I. (1984). Educational diagnostic assessment. *Journal of Educational Measurement*, *21*(2), 175-189. <https://doi.org/10.1111/j.1745-3984.1984.tb00228.x>

Bentler, P. M., & Dudgeon, P. (1996). Covariance Structure Analysis: Statistical Practice, Theory, and Directions. *Annual Review of Psychology*, *47*(1), 563-592. <https://doi.org/10.1146/annurev.psych.47.1.563>

Bloom, B. S., Madaus, G. F., & Hastings, J. T. (1981). *Evaluation to improve learning*. McGraw-Hill.

Bonner, S. M. (2016). Teachers’ Perceptions about Assessment: Competing Narratives. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of Human and Social Conditions in Assessment* (pp. 21-39). Routledge.

Brown, G. T. L. (2003). *Teachers' conceptions of assessment inventory-Abridged (TCoA Version IIIA) [Measurement instrument]*. <https://doi.org/10.17608/k6.auckland.3199543.v1>

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. <https://doi.org/10.1080/0969594042000304609>

Brown, G. T. L. (2006). Teachers' conceptions of assessment: Validation of an abridged version. *Psychological Reports*, *99*(1), 166-170. <https://doi.org/10.2466/pr0.99.1.166-170>

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

Brown, G. T. L. (2011). Teachers' conceptions of assessment: Comparing primary and secondary teachers in New Zealand. *Assessment Matters*, *3*, 45-70. <https://doi.org/10.18296/am.0097>

Brown, G. T. L. (2013). asTTle—A national testing system for formative assessment: How the national testing policy ended up helping schools and teachers. In M. K. Lai & S. Kushner (Eds.), *A National Developmental and Negotiated Approach to School and Curriculum Evaluation* (pp. 39-56). Emerald Group Publishing. <https://doi.org/10.1108/S1474-7863(2013)0000014003>

Brown, G. T. L. (2017). Improvement and accountability functions of assessment: Impact on teachers’ thinking and action. In M. A. Peters (Ed.), *Encyclopedia of Educational Philosophy and Theory* (pp. 1109–1114). Singapore: Springer Singapore. <https://doi.org/10.1007/978-981-287-532-7_391-1>

Brown, G. T. L. (2019). Technologies and infrastructure: costs and obstacles in developing large-scale computer–based testing. *Education Inquiry*, *10*(1), 4-20. <https://doi.org/10.1080/20004508.2018.1529528>

Brown, G. T. L., & Harris, L. R. (2009). Unintended consequences of using tests to improve learning: How improvement-oriented resources heighten conceptions of assessment as school accountability. *Journal of MultiDisciplinary Evaluation*, *6*(12), 68-91. <https://doi.org/10.56645/jmde.v6i12.236>

Brown, G. T. L., & Shulruf, B. (2023). Response Option Design in Surveys. In L. R. Ford & T. A. Scandura (Eds.), *The SAGE Handbook of Survey Development and Application* (pp. 120-132). Sage.

Brown, G. T. L., Gebril, A., & Michaelides, M. P. (2019). Teachers' Conceptions of Assessment: A Global Phenomenon or a Global Localism. *Frontiers in Education*, *4*(16). <https://doi.org/10.3389/feduc.2019.00016>

Brown, G. T. L., Harris, L. R., O'Quin, C., & Lane, K. E. (2017). Using multi-group confirmatory factor analysis to evaluate cross-cultural research: identifying and understanding non-invariance. *International Journal of Research & Method in Education*, *40*(1), 66-90. <https://doi.org/10.1080/1743727X.2015.1070823>

Brown, G. T. L., Hui, S. K. F., Yu, F. W. M., & Kennedy, K. J. (2011). Teachers' conceptions of assessment in Chinese contexts: A tripartite model of accountability, improvement, and irrelevance. *International Journal of Educational Research*, *50*(5-6), 307-320. <https://doi.org/10.1016/j.ijer.2011.10.003>

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

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

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

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge.

Courtney, M. G. R., & Chang, K. C. (2018). Dealing with non-normality: An introduction and step-by-step guide using R. *Teaching. Statistics, 40*(2), 51–59. https://doi.org/10.1111/ test.12154

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, *39*(1), 1-38. <https://doi.org/10.1111/j.2517-6161.1977.tb01600.x>

DiStefano, C., Zhu, M., & Mîndrilă, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation*, *14*(20). <https://doi.org/10.7275/da8t-4g52>

EVS/WVS. (2022). European Values Study and World Values Survey: Joint EVS/WVS 2017-2022 Dataset (Joint EVS/WVS). *GESIS Data Archive*, Cologne. ZA7505. Dataset Version 4.0.0,. https://doi.org/ [10.4232/1.14023](http://dx.doi.org/10.4232/1.14023)

Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research*, *42*(3), 509–529. <https://doi.org/10.1080/00273170701382864>

Finney, S. J., & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. In G. R. Hancock & R. D. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 269-314). Information Age Publishing.

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

Fulmer, G. W., Lee, I. C. H., & Tan, K. H. K. (2015). Multi-level model of contextual factors and teachers’ assessment practices: an integrative review of research. *Assessment in Education: Principles, Policy & Practice*, *22*(4), 475-494. <https://doi.org/doi:10.1080/0969594X.2015.1017445>

Hancock, G. R., & Mueller, R. O. (2001). Rethinking Construct Reliability Within Latent Variable Systems. In R. Cudeck, S. Du Toit, & D. Sörbom (Eds.), *Structural Equation Modeling: Present and Future - A Festschrift in Honor of Karl Jöreskog* (pp. 195-216). Scientific Software International Inc.

Harris, L. R., & Brown, G. T. L. (2009). The complexity of teachers’ conceptions of assessment: tensions between the needs of schools and students. *Assessment in Education: Principles, Policy & Practice*, *16*(3), 365-381. <https://doi.org/10.1080/09695940903319745>

Helgøy, I., & Homme, A. (2007). Towards a New Professionalism in School? A Comparative Study of Teacher Autonomy in Norway and Sweden. *European Educational Research Journal*, *6*(3), 232-249. https://doi.org/10.2304/eerj.2007.6.3.232

Hofstede, G., Hofstede, G. J., & Minkov, M. (2010). *Cultures and organizations: software of the mind* (3. ed. ed.). New York : McGraw-Hill.

Hofverberg, A., & Winberg, M. (2020). Challenging the universality of achievement goal models: A comparison of two culturally distinct countries. *Scandinavian Journal of Educational Research, 64*(3), 333-354. <https://doi.org/10.1080/00313831.2018.1544170>

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1-55. <https://doi.org/10.1080/10705519909540118>

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*(351a), 631-639. <https://doi.org/10.1080/01621459.1975.10482485>

Kim, H.-Y. (2013). Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. *Restorative Dentistry & Endodontics, 38*(1), 52–54. https://doi.org/10.5395/rde.2013.38.1.52

Lai, M. K., & Schildkamp, K. (2016). In-service teacher professional learning: Use of assessment in data-based decision-making. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of Human and Social Conditions in Assessment* (pp. 77-94). Routledge.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American statistical Association.* 83, 1198–1202. https://doi.org/10.2307/2290157

Maydeu-Olivares, A., Shi, D., & Rosseel, Y. (2018). Assessing fit in structural equation models: A Monte-Carlo evaluation of RMSEA Versus SRMR confidence intervals and tests of close fit. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(3), 389-402. https://doi.org/10.1080/10705511.2017.1389611

McArdle, J. J. (2007). Five steps in the structural factor analysis of longitudinal data. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (pp. 99-130). LEA.

Nichols, S. L., & Harris, L. R. (2016). Accountability assessment’s effects on teachers and schools. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of Human and Social Conditions in Assessment* (pp. 40-56). Routledge.

Revelle, W., & Rocklin, T. (1979). Very simple structure: An alternative procedure For estimating the optimal number of interpretable factors. *Multivariate Behavioral Research*, *14*(4), 403-414. <https://doi.org/10.1207/s15327906mbr1404_2>

SFS 2010:800. *Skollag [The Education Act].* Stockholm: Utbildningsdepartementet [Ministry of Education].

Skolverket [National Agency for Education]. (2020). Att planera, bedöma och ge återkoppling: stöd för undervisning [To plan, assess, and give feedback: Support for teaching]. Stockholm: Skolverket

Skolverket [National Agency for Education]. (2022). *Läroplan för grundskolan, förskoleklassen och fritidshemmet [Curriculum for compulsory school, pre-school class and school age educare], Lgr 22*. Stockholm: Skolverket.

Tong, A., Sainsbury, P., & Craig, J. (2007). Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *International journal for quality in health care, 19*(6), 349–357. https://doi.org/10.1093/intqhc/mzm042

Wheaton, B., Muthén, B., Alwin, D. F., & Summers, G. F. (1977). Assessing reliability and stability in panel models. *Sociological methodology*, *8*, 84-136. <https://doi.org/10.2307/270754>

Table 1. Demographic characteristics: Matched for two administrations of the TCoA.

|  |  |  |
| --- | --- | --- |
| Demographic | Frequency *n* | % |
| Sex |  |  |
| Women | 202 | 79 |
| Men | 34 | 13 |
| Missing | 21 | 8 |
| Degree status |  |  |
| Held | 198 | 77 |
| Missing or no | 59 | 23 |
| Teaching experience |  |  |
| Less than 2 years | 15 | 6 |
| 2-5 years | 28 | 11 |
| 6-10 years | 27 | 11 |
| >10 years | 163 | 63 |
| Missing | 24 | 9 |
| School stage |  |  |
| Y1-3 | 77 | 30 |
| Y4-6 | 43 | 17 |
| Y7-9 | 58 | 23 |
| Missing | 79 | 30 |

Table 2. Swedish TCoA factor structure, item loadings, item wording in Swedish and English, and factor robustness by time

| Predictor | Loading | | Coeff. H | | *M* (*SD*) | |
| --- | --- | --- | --- | --- | --- | --- |
| Dependent item / factor | 2021 | 2023 | 2021 | 2023 | 2021 | 2023 |
| Describe |  |  | .62 | .65 | 4.04 (.82) | 3.94 (.86) |
| Dia1 Bedömning är ett sätt att fastställa hur mycket elever har lärt sig av undervisningen. [Assessment is a way to determine how much students have learned from teaching] | 0.55 | 0.57 |  |  |  |  |
| Dia2 Bedömning fastställer vad elever har lärt sig.[Assessment establishes what students have learned] | 0.59 | 0.62 |  |  |  |  |
| Dia3 Bedömningarna som görs i skolan mäter elevers förmåga till analytiskt och reflekterande tänkande. [Assessment measures students’ higher order thinking skills] | 0.63 | 0.65 |  |  |  |  |
| Improve Student Learning |  |  | .79 | .74 | 4.15 (.97) | 3.73 (.88) |
| si1 Bedömningarna som görs i skolan ger återkoppling till elever om deras prestationer. [Assessment provides feedback to students about their performance] | 0.61 | 0.62 |  |  |  |  |
| si2 Bedömningarna som görs i skolan ger elever återkoppling på deras lärandebehov. [Assessment feedbacks to students their learning needs] | 0.69 | 0.67 |  |  |  |  |
| si3 Bedömningarna som görs i skolan hjälper elever att förbättra sitt lärande. [Assessment helps students improve their learning] | 0.83 | 0.76 |  |  |  |  |
| Improve Teaching |  |  | .65 | .71 | 4.09 (.96) | 4.12 (1.04) |
| ti1 Bedömning är en integrerad del av undervisning. [Assessment is integrated with teaching practice] | 0.52 | 0.61 |  |  |  |  |
| ti2 Lärare modifierar pågående undervisning av elever baserat på bedömningsinformation. [Assessment information modifies ongoing teaching of students] | 0.59 | 0.64 |  |  |  |  |
| ti3 Bedömning möjliggör att olika elever kan få olika undervisning. [Assessment allows different students to get different instruction] | 0.70 | 0.74 |  |  |  |  |
| Valid |  |  | .73 | .78 | 3.46 (.82) | 3.73 (.88) |
| Rel1 Bedömningsresultat är tillförlitliga. [Assessment results are trustworthy] | 0.57 | 0.73 |  |  |  |  |
| Rel2 Bedömningsresultat blir desamma oavsett vem som bedömer. [Assessment results are consistent] | 0.48 | 0.53 |  |  |  |  |
| Rel3 Bedömningsresultat är pålitliga [Assessment results can be depended on] | 0.81 | 0.82 |  |  |  |  |
| School Quality |  |  | .80 | .89 | 3.37 (.98) | 3.37 (1.08) |
| sq1 Bedömning ger information om hur bra det går för skolor. [Assessment provides information on how well schools are doing] | 0.58 | 0.70 |  |  |  |  |
| sq2 Bedömningsresultat är en rättvisande indikator på en skolas kvalitet. [Assessment is an accurate indicator of a school’s quality] | 0.71 | 0.78 |  |  |  |  |
| sq3 Bedömning är ett bra sätt att utvärdera en skola. [Assessment is a good way to evaluate a school] | 0.85 | 0.92 |  |  |  |  |
| Irrelevance |  |  | .59 | .74 | 2.91 (.87) | 2.91 (1.00) |
| Ig1 Lärare genomför bedömningar men använder resultaten i låg utsträckning. [Teachers conduct assessments but make little use of the results] | 0.46 | 0.58 |  |  |  |  |
| Ig2 Bedömning är orättvist mot eleverna. [Assessment is unfair to students] | 0.60 | 0.69 |  |  |  |  |
| Ig3 Bedömningsresultat arkiveras och ignoreras. [Assessment results are filed & ignored] | 0.51 | 0.68 |  |  |  |  |
| Ir4 Bedömning stör undervisningen. [Assessment interferes with teaching] | 0.47 | 0.58 |  |  |  |  |
| Improvement |  |  | 1.00 | .99 |  |  |
| Improve Student Learning | 1.00 | 0.99 |  |  |  |  |
| Valid | 0.74 | 0.79 |  |  |  |  |
| Describe | 0.87 | 0.97 |  |  |  |  |
| Improve Teaching | 0.85 | 0.85 |  |  |  |  |

Note. Loadings are standardised beta weights; values constrained to equivalent measurement weights.

Table 3. Factor inter-correlations by data wave

|  |  |  |  |
| --- | --- | --- | --- |
| TCoA Scale | I. | II. | III. |
| I. Improvement | — | .52\*\*\* | -.61\*\*\* |
| II. School Quality | .47\*\*\* | — | -.11ns |
| III. Irrelevance | -.60\*\*\* | .02ns | — |

Note. 2021 values below diagonal; 2023 values above diagonal; values constrained to equivalent measurement weights; ns = not significant; \*\*\* = *p*<.001.

Table 4. Statistically Significant Tukey HSD results by Teaching Level

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Teaching Level | | |
| Time | Scale | Y7-9 (*n* = 48) | Y4-6 (*n* = 40) | Y1-3 (*n* = 75) |
| 1 | Assessment is Valid | 3.68 | 3.26 | 3.41 |
| 2 | Assessment Helps Students Improve | 4.12 | 3.66 | 3.64 |
| 2 | Assessment Provides Diagnostic Description | 4.24 | 3.83 | 3.77 |
| 2 | Assessment is Irrelevant | 2.58 | 3.06 | 3.20 |

Figure 1. Swedish Teachers’ Conceptions of Assessment: Invariance at measurement weights only

A diagram of a model

Description automatically generatedNote. All values are standardised; error terms omitted for simplicity; bold = 2021; italic = 2023

1. https://autopsych.shinyapps.io/normalr/ [↑](#footnote-ref-1)