Assessing Measurement Invariance of the Teachers' Perceptions of Grading Practices Scale across Cultures

Xing Liu
Assistant Professor
Education Department
Eastern Connecticut State University
83 Windham Street,
Willimantic, CT, 06226, U.S.A.

Abstract

In a cross-cultural study, it is crucial to understand whether items and the factorial structure of a survey instrument are equivalent across different cultural groups, because items might mean differently to different population groups, and thus the factorial structure of the measurement instrument might not hold across groups. The purpose of this study was to investigate the measurement invariance of the Teachers' Perceptions of Grading Practices Scale (TPGP) across the U.S. and China using structural equation modeling. In particular, this study was designed to examine whether items and the factorial structure of the TPGP scale were equivalent across the two countries, and identify nonequivalent items if this instrument was a partially invariant measurement instrument.

Key words: Measurement Invariance, structural equation modeling, multi-group CFA, Teachers' Perceptions of Grading Practices (TPGP), Cross-Culture

1. Introduction

The effectiveness of classroom assessment and grading practices has become an increasingly important research topic in education (Bonesronning, 1999, 2004; Brookhart, 1993, 1994; McMillan & Lawson, 2001; McMillan, Myran, & Workman, 2002; McMillan & Nash, 2000). Numerous studies have been conducted regarding factors affecting teachers' grading practices (Brookhart, 1993, 1994; McMillan & Lawson, 2001; McMillan, Myran, & Workman, 2002; McMunn, Schenck, & McColskey, 2003; Stiggins, Frisbie, & Griswold, 1989). A self-report survey instrument, Teachers' Perceptions of Grading Practices Scale (TPGP), was recently developed to measure teachers' perceptions (Liu, 2004; Liu, O'Connell, & McCoach, 2006). The initial validation of this instrument appeared to have sound psychometric properties and good reliability. However, in a cross-cultural study, it was crucial to understand whether items and factorial structure of a survey instrument were equivalent across different cultural groups, because items might mean differently to different population groups, and thus the factorial structure of the measurement instrument might not hold across groups. When a measurement instrument was not equivalent in a cross-cultural study, the validity of research findings would be problematic and need further investigation.

The purpose of this study was to investigate the measurement invariance of the Teachers' Perceptions of Grading Practices Scale (TPGP) across the U.S. and China using structural equation modeling. In particular, this study was designed to examine whether items and the factorial structure of the TPGP scale were equivalent across the two countries, and to determine those nonequivalent items if the instrument was a partially invariant measurement instrument. This study could help researchers and school administrators to develop and validate an instrument to understand teachers' perceptions of grading practices across cultures. It would also provide empirical evidence for researchers to deal with partial measurement invariance and how to identify nonequivalent items of an instrument in cross-cultural research.

2. Theoretical Framework

Multiple-group confirmatory factor analysis was an appealing approach to examine whether items and the factorial structure of a measurement instrument were equivalent across different groups (Brown, 2006; Byrne, 2001; Joreskog, 1971).

According to Brown (2006), one advantage of multiple-group confirmatory factor analysis was that all aspects of measurement invariance (i.e. factor loadings, intercepts, error variances) and structural invariance (i.e. factor variance-covariance structure and factor means) could be examined across different populations. The most commonly tested parameters in multiple-group analysis are factor loadings, factor variances and covariances, and structural regression paths (Byrne, 2001). As suggested by Arbuckle (2005) and Brown (2006), the orderly sequence for examining measurement and structural invariance is as follows: (1) Preliminary separate single-group CFA analyses across groups; (2) a baseline multiple-group model analysis with no equality constraints imposed; (3) a model with equality constraints across groups specified for measurement weights (factor loadings); (4) a model with equality constraints across groups specified for measurement intercepts; (5) a model with equality constraints across groups specified for factor variances and covariances; and (6) a model with equality constraints across groups specified for measurement residuals .

This process of model fitting from steps 2-6 yielded a nested hierarchy of models in which each model contained all the constraints of the prior model, and thus, each was nested within its earlier models. Chi-square difference tests were used to test whether the equality constraints were upheld. A non significant chi-square difference test suggests that the equality constraints imposed on these parameters (e.g. factor loadings) are tenable (Byrne, 1989). Among these tests for the above equality constraints, the test of equal factor loadings of items is critical (Brown, 2006). If the assumption of equal factor loading is upheld, it suggests that these items convey the same meaning for samples across different groups, and then the latent construct underlying these items are comparable across groups. However, Brown (2006) argued that if this assumption is violated, it is not appropriate to proceed to conduct other tests of more restrictive constraints (e.g. equal factor variances and covariances, or factor means). "Group comparisons of factor variances and covariances are meaningful only when the factor loadings are invariant" (Brown, 2006, p. 269).

To deal with this issue, Byrne, Shavelson, and Muthen (1989) demonstrated how to examine measurement and structural invariance in the context of partial measurement invariance. According to Byrne et al. (1989) and Brown (2006), partial measurement invariance means that some but not all of the measurement parameters (e.g. factor loadings) are equivalent across groups in a multiple-group CFA model. If factor loadings of some items in an instrument were identified as noninvariant, this instrument is the one with partial measurement invariance. To identify factor loadings of which items are equivalent and which are nonequivalent across groups, chi-square difference tests are recommended by Byrne et al. (1989) on an item-by-item basis. First, a model is fit by placing equality constraints on all the factor loadings, then, a less restrictive model is fitted by relaxing the equality constraint of the regression weights of the item of interest. A non-significant chi-square difference test indicates that the factor loading of that item is not statistically different across groups. This process can be repeated item by item until all of the nonequivalent items are identified.

According to Yuan and Bentler (2007), in real world data analysis, it is difficult to achieve accurate assessment of invariance on parameters across groups. Byrne and Watkins (2003) conducted a multi-group confirmatory factor analysis to test the equivalence of the Self Description Questionnaire I (SDQ-I), a well-known measurement instrument used in cross-cultural research, across two culturally diverse groups for Australian and Nigerian adolescents. They found that the factorial structure of the instrument was similarly specified and well-fit for each separate cultural group, but they also found evidence of both measurement and structural non-invariance across two groups. To deal with this issue, the researchers conducted more detailed analyses to investigate item invariance and identify nonequivalent parameters (e.g., factor loadings) across Australia and Nigerian adolescents in the context of partial measurement invariance. On the basis of the finding of Byrne and Watkins, this researcher decided that a similar approach should be used in this study to determine the invariance of factor loadings for the TPGP instrument across the U.S. and China, and identify which items were nonequivalent.

In a confirmatory factor analysis, an initial model can be respecified in order to improve its goodness of fit, parsimony and interpretability of the model (Brown, 2006). Model respecification is based on modification indices (empirical evidence) and substantive justification (theoretical evidence). Trimming off indicators with low loadings and correlating errors of indicators are two general ways of model respecification (Kline, 2005). Correlated errors are specified when some of the covariance across two indicators is not explained by the latent construct (Brown, 2006). Although correlated error can be specified according to modification indices, they need to be supported by a theoretical rationale.

In some situations, according to Brown (2006), in the analysis of survey items, item errors may be correlated when these items are "very similarly worded, reverse-worded, or differentially prone to social desirability, and so forth" (p. 181). In this study, when correlated errors were specified on the CFA models, theoretical evidence for why these errors were correlated was provided. Also, fit indices of both the original model and respecified model were examined.

3. Data and Methods

3.1 Instrumentation

An instrument, the *Teachers' Perceptions of Grading Practices* (TPGP), was developed to assess teachers' perceptions (Liu, 2004; Liu, O'Connell, & McCoach, 2006); this instrument measuring teachers' perceptions of grading practices has six sections. Table 1 provides sections and items for the final survey. The survey instrument was designed in both English and Chinese versions so that the teachers in China took the Chinese version and those in the U.S. took the English one. To ensure the translation validity, a back translation was conducted after the English-version survey was translated into Chinese. To complete the survey, participants were asked to circle or click on their answer to each item with responses ranging from strongly disagree to strongly agree based on 5-point Likert rating scale (1 = Strongly Disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, and 5 = Strongly Agree).

3.2 Sample and Data Collection

The target populations of the study are middle and high school teachers in the U.S. and China. The sample was selected from a state in the Northeastern United States, and a city in a province in China. The targeted sample size was 150 from each country, because previous research recommended the sample size under structural equation modeling (SEM) to be at least 100-150 cases (Klem, 2000). Both random sampling and convenience sampling techniques were used in the study. After human subject protocol approval was obtained through the University of Connecticut (UConn), the data collection began in both countries.

In the U.S., self-report web-based surveys were used to gather the data. Participants were asked to respond to the survey items by following the directions online. Responses were anonymous; respondents were not required to provide names that could be linked to their responses. To increase the response rate of the online surveys in the U.S., two i-pods were used as incentives for a raffle, since previous research found that using an incentive could increase response rates to the online survey (Dommeyer, Baum, & Hanna, 2004). The participants who received emails were asked to enter their email addresses at the end of the online survey if they were willing to join in the raffle. Those participants who received requests through the regular mail were asked to enter codes (numbers) which were assigned and mailed to each of them in the letter, or enter their email addresses. These codes (numbers) were used for raffle purposes only, since the raffle needed to link the name with the corresponding code.

Ensuring confidentiality has been found to increase response rates in survey research (Asch, Jedriziewski, & Christakis, 1997). To ensure confidentiality, two separate files were programmed into the on-line survey, one to collect the email addresses or codes so we knew who completed the survey to enter the raffle, and one to collect the actual data, which was not linked back to the email address or codes. In the cleaned final data set, no information on the teachers' name, email addresses and codes are identified. All survey data were entered into a secure, restricted database. All information was kept confidential and only researchers could have access to the data, which were kept in a locked cabinet in the researcher's office. When reporting the results, no individual names or school district could be identified. The results are reported only on the group level. The purpose of using the email addresses or codes and these procedures to ensure confidentiality were fully explained to the participants in the emails and letters. A total of 609 middle and high school teachers were contacted by email or mail, and a total of 122 secondary school teachers responded to the online survey, with a response rate of 20%.

In China, surveys were sent out to teachers from urban middle and high schools in Taizhou City, Jiangsu Province, using a cluster sampling technique. Instead of using online surveys, the method of dropping-off/picking-up was used to gather the data in China. As has been supported through previous research, there was no expectation of differences in data and response quality between web-based and non-web-based surveys of data collection (Fiala, 2005). For the China-sample, first, five schools (three middle schools and two high schools) were randomly selected from 12 urban secondary schools in the central Taizhou City using simple random sampling; then paper surveys were sent to all teachers' mailboxes within these selected schools, which had a population of 400 teachers. For each school, one correspondent of the school administration was appointed to drop-off and collect the surveys.

The surveys were collected and returned anonymously. A total of 167 teachers responded and returned the surveys, with a response rate of 42%. Demographic comparisons are provided in Table 1.

3.3 Data analysis

A preliminary single-group confirmatory factor analysis was conducted to examine the factorial structure of the TPGP instrument for the full-sample data from both countries. Then, two separate single-group confirmatory factor analyses were conducted for each country. Finally, a multi-group factor analysis was conducted simultaneously across two countries to examine the construct validity of the six-factor and 40-item TPGP instrument across samples from the U.S. and China. In an effort to identify factor loadings of those items found to be equivalent and those found to be nonequivalent across countries, chi-square difference tests were conducted on an item-by-item basis within the context of partial measurement invariance. AMOS (19.0) was used for both single-group confirmatory factor analysis and multi-group confirmatory factor analysis.

4. Results

Maximum likelihood estimation was used to estimate the parameters. Multiple indices such as chi-square test, relative χ^2/df, Comparative Fit Index (CFI), and Root Mean Square Error of Approximation (RMSEA) were used to evaluate the model fit. Normally a non-significant chi-square result indicates a good model fit. However, the Chi-square test is not a satisfactory test of model fit considering its dependency on sample size (Bentler & Bonett, 1980; Byrne, 1994). Therefore, several additional fit statistics were considered together with the Chi-square test. As a rule of thumb, values of relative χ^2/df less than two or three indicate a good model fit, values of RMSEA less than .08 indicate a reasonable fit, and values of CFI larger than .90 indicate an acceptable fit (Hu & Bentler, 1999). For comparison of model improvement of fit among nested models, the χ^2 difference test was used.

4.1 *Preliminary Single-group Confirmatory Factor Analysis for the Full-sample Data*

First, a preliminary single-group confirmatory factor analysis was conducted using the full-sample data (n = 389) (unmodified hypothesized model, in Table 2). In the unmodified hypothesized model under test, each item was specified as an indicator for only one factor, and no errors were correlated; items 39 and 40 were reverse scored so that the directions of all items were consistent (Table 3). The fit indices for the hypothesized six-factor model with 40 items were as follows: $\chi^2 = 1562.67$, df = 687, $p < .001$. CFI = .80, RMSEA = .067 (90% Confidence interval of .062 to .071), and $\chi^2/df = 2.277$. The model was re-specified after examining the standardized regression weights (factor loadings), the squared multiple correlations of the items, and the modification indices. Based on modification indices, correlated errors between item eight and item nine, item 12 and item 15, 17 and item 18, item 20 and item 21, and item 26 and item 27 were added to the fitted model. In addition to suggestions by modification indices, these correlated errors were also supported by a theoretical rationale. For instance, item eight "Grading can help me improve instruction" and item nine "Grading can encourage good work by students" had a stronger connection since good work by students could be encouraged by a good quality of instruction of a teacher. In addition, according to Brown (2006), when items in a survey were very similarly worded, their errors might be correlated. Since some of the items above were similarly worded (e.g., items eight and nine, items 17 and 18, items 20 and 21, and items 26 and 27), this explanation provided evidence for the correlated errors of these items.

This re-specified model was named Modified Hypothesized Model 1 (Table 4.1a). Fit indices showed that the re-specified model resulted in a significant improvement of fit, compared to the originally unmodified model, $\chi^2_{diff}(5) = 82.82$, $p<.001$. To improve the fit indices of CFI, this model was specified again. Seven items (items 29, 32, 33, 34, 36, 39, and 40) with low standardized regression weights (less than .50) and squared multiple correlations (less than .15) were trimmed off the model. Two additional correlated errors between item five and item 13 and between item 24 and item 26 were added to model. Item five "Grading practices are important measures of student achievement" and item 13 "Grading provides information about student achievement" had a stronger connection because both items focused on the relations between grading and student achievement. Item 24 "Grades are based on students' problem solving ability" and item 26 "Grades are based on students' independent thinking ability" had a stronger connection because these two abilities were both needed by students, and these abilities played an important role in teachers' grading decision. This re-specified model was named Modified Hypothesized Model 2.

The fit indices for this model were $\chi^2 = 949.03$, df = 473, $p < .001$. CFI = .88, RMSEA = .059 (90% Confidence interval of .054 to .065), and $\chi^2/df = 2.006$. The fit indices of RMSEA and relative chi-square suggested an acceptable model fit. CFI was close to the cut-off point of acceptable fit (.90). Overall, the hypothesized six-factor model exhibited borderline fit; not all fit indices were strong.

4.2 Preliminary Single-group Analyses for U.S.-sample Data

Next, the above three single-group CFA models were fitted using the U.S.-sample data only (n = 122). Table 4 presents summary fit indices from confirmatory factor analysis for the US-sample data. The fit indices for the final re-specified model were $\chi^2 = 684.65$, df = 473, $p < .001$. CFI = .87, RMSEA = .061 (90% Confidence interval of .051 to .071), and $\chi^2/df = 1.447$. These fit indices suggested that the model exhibited borderline fit.

4.3 Preliminary Single-group Analyses for China-sample Data

The above three single-group CFA models were fit using the China-sample data only (n = 167). Table 5 presents summary fit indices from confirmatory factor analysis for the China-sample data. The fit indices for the final re-specified model were $\chi^2 = 893.18$, df = 473, $p < .001$. CFI = .85, RMSEA = .073 (90% Confidence interval of .066 to .080), and $\chi^2/df = 2.008$. The modified model one resulted in a significant improvement of fit, compared to the originally unmodified model, $\chi^2_{diff}(5) = 47.84$, $p<.001$).

4.4 Multi-group Confirmatory Factor Analysis

The results of the above single-group CFA model using the full-sample data and the two separate single-group CFA models across each country indicated that the factorial structure of the TPGP instrument exhibited borderline fit overall, and this structure demonstrated a similar pattern within each of two countries. To test whether the factorial structure of the TPGP instrument was equivalent across samples of two countries, a multi-group CFA was conducted following the procedures below:

1. A baseline model was constructed with no equality constraints imposed;
2. Models were fit with equality constraints across countries specified for measurement weights (factor loadings), measurement intercepts, structural covariances (factor variances and covariances), and measurement residuals (variances and covariances of residual variables), respectively.
3. This process of model fitting yielded a nested hierarchy of models in which each model contained all the constraints of the prior model, and thus, each was nested within its earlier models. A chi-square difference test was used to test whether the equality constraints were upheld.

Table 6 presents summary fit indices of five nested models for the multi-group confirmatory factor analysis. The unconstrained model was the baseline model, which relaxed all equality constraints. This model tested the factorial structure of the instrument across two countries simultaneously with no cross-group constraints imposed. The fit indices for the baseline model were $\chi^2 = 1577.87$, df = 946, p = .000. CFI = .86, RMSEA = .048 (90% confidence interval of .044 to .052), and $\chi^2/df = 1.668$. These indices indicated that the hypothesized six-factor model of TPGP instrument exhibited acceptable fit across samples of two countries.

The measurement weights model tested the invariance of factor loadings across countries by placing equality constraints on these parameters. The fit indices for this model were $\chi^2 = 1668.25$, df = 973, p = .000. CFI = .84, RMSEA = .050 (90% confidence interval of .046 to .054), and $\chi^2/df = 1.715$. Since the measurement weights model was nested within the unconstrained model, the chi-square difference test, $\chi^2_{diff}(27) = 90.38$, $p<.001$, indicated that some equality constraints of factor loadings did not hold across two countries. A detailed exploration of which loadings were different across the two groups is provided in Analysis Four (see partial measurement invariance below)

The measurement intercepts model placed equality constraints across groups on intercepts in the equations for predicting items, in addition to equality constraints on factor loadings. Compared to the measurement weights model, the chi-square difference test, $\chi^2_{diff}(33) = 384.41$, $p<.001$, indicating that some equality constraints of intercepts did not hold across the two countries. Next, when factor variances and covariances were constrained equally across countries, compared to the measurement intercepts model, the chi-square difference test, $\chi^2_{diff}(21) = 71.79$, $p<.001$, indicating that the matrices of factor variances and covariances were not equal across the two countries. In the measurement residual model, all parameters were specified equally across countries. The chi-square difference test again yielded a statistically significant value of 263.57 with 40 degrees of freedom at the .01 level.

These findings suggested that the equality constraints of factor loadings, intercepts, factor variances and covariances, and error covariances were not upheld across the two countries. That is, the assumption of an equivalent factor structure was not supported across the two countries.

4.5 Partial Measurement Invariance

The results of the measurement weights model analysis above indicated that some equality constraints of factor loadings did not hold across the two countries. In an effort to identify factor loadings of which items were equivalent and which were nonequivalent across countries, chi-square difference tests were conducted on an item-by-item basis in the context of partial measurement invariance where equality constraints were imposed on some but not all of the factor loadings (Byrne, Shavelson, & Muthen, 1989). In the context of the partial measurement model, a model was fitted first by placing equality constraints on all the factor loadings (See measurement weights model in Table 6); then, a less restrictive model was fit by relaxing the equality constraint of the regression weights (factor loadings) of the item of interest.

A chi-square difference test was conducted between the less restrictive model and the measurement weights model to investigate whether the factor loading of that item was invariant across the two countries. A non-significant chi-square difference test indicated that the factor loading of that item was not statistically different across the two countries. This process was repeated item by item until all of the nonequivalent items were identified. In order to identify whether marker indicators (items 1, 8, 17, 23, 30 and 35) were equivalent across the two countries, following the procedure by Brown (2006), chi-square difference tests were conducted after re-running multi-group CFA with different marker indicators. When testing the marker indicator itself, another marker indicator needs to be selected in the model. For instance, when the marker indicator, item one was tested whether it was equivalent across groups, the equality constraint of regression weights of this item was relaxed, and an equivalent item, item two was chosen as a marker indicator. A chi-square difference test was conducted for each marker item.

Results of chi-square difference tests indicated that the factor loadings of items 6, 9, 12, 13, 20, and 23 were nonequivalent (item 26 was treated as an equivalent item, since the significant level of the chi-square difference test for this item was close to .05), and the factor loadings of other items were invariant across countries. Table 7 displays the results of chi-square difference tests for equivalent factor loadings and nonequivalent factor loadings of items across countries. Table 8 displays factor loading of items across the two countries. Therefore, the research suggested that the TPGP instrument was a partially invariant measurement instrument across the two countries because the factor loadings of some items were not equivalent across the two samples. Although evidence of nonequivalence could be determined by multiple-group CFA analysis in the context of partial measurement invariance, the technique itself could not explain the reasons for nonequivalence. When nonequivalent items were identified, some possible reasons might be different interpretation and different social desirability across cultures (Byrne & Watkins, 2003).

5. Conclusions and Discussion

In this study the results of multi-group CFA analyses suggested that some equality constraints of factor loadings, intercepts, factor variances and covariances, and error covariances were not upheld across two countries. Chi-square difference tests were conducted to determine the invariance of factor loadings for the TPGP instrument on an item-by-item basis in the context of partial measurement invariance. Results of chi-square difference tests indicated that the factor loadings of items six, nine, 12, 13, 20, and 23 were nonequivalent, and the factor loadings of other items were invariant across countries. These findings suggested that the TPGP instrument was a partially invariant measurement instrument across the two countries, because the pattern coefficients were not equivalent across the two samples. These six nonequivalent items (items six, nine, 12, 13, 20, and 23) covered the topics of the importance of grading (Item six), the usefulness of grading (items nine, 12, and 13), student effort (item 20), and student ability (item 23). The reason for nonequivalence in factor loadings for these items might be due to translation, different interpretations and different social desirability across cultures (Byrne, & Watkins, 2003). Although a back translation was conducted after the English-version survey was translated into Chinese, and two bilingual experts reviewed both versions of the survey, translation still might be an issue for certain items. Furthermore, these items might mean different things for teachers in the U.S. and China, and in a certain cultural context, teachers responded to some particular items homogeneously due to social desirability.

This study provided empirical evidence of how to deal with partial measurement invariance and how to identify nonequivalent items of an instrument in cross-cultural research.

References

Arbuckle, J. L. (2010). *Amos 19.0 user's guide*. Chicago, IL: SPSS Inc.

Asch, D., Jedriziewski, M., & Christakis, N. (1997). Response rates to mail survey published in medical journal. *Journal of Clinical Epidemiology, 50* (10), 1129-1136.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588-606.

Bonesronning, H. (1998). The variation in teachers' grading practices: causes and consequences. *Economics of Education Review, 18*, 89-105.

Bonesronning, H. (2004). Do the teachers' grading practices affect student achievement? *Education Economics, 12*, 151-167.

Brookhart, S. M. (1993). Teachers' grading practices: Meaning and values. *Journal of Educational Measurement, 30*, 123-142.

Brookhart, S. M. (1994). Teacher's grading: Practice and theory. *Applied Measurement in Education, 7*, 279-301.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: The Guilford Press.

Byrne, B. M. (1994). Testing for the factorial validity, replication, and invariance of a measuring instrument: A paradigmatic application based on the Maslach Burnout Inventory. *Multivariate Behavioral Research, 29*, 289-311.

Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum.

Byrne, B. M., Shavelson, R. J., & Muthen, B. (1989). Testing for the equivalentce of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*, 456-466.

Byrne, B. M., & Watkins, D. (2003). The issue of measurement invariance revisited. *Journal of Cross-cultural Psychology, 34 (2),* 155-175.

Dommeyer, C. J., Baum, P., & Hanna, R. W. (2004). Gathering faculty teaching evaluations by in-class and online surveys: Their effects on response rates and evaluations. *Assessment & Evaluation in High Education. 29* (5) 611-623.

Fiala, A. K. (2005). A comparison of characteristics and responses for non-responding and responding certified athletic trainers to mail and web-based surveys on continuing education units. *Dissertation Abstract International, 65* (12), 4460. (UMI No. 3156389)

Gullickson. A. R. (1985). Student evaluation techniques and their relationship to grade and curriculum. *Journal of Educational Research. 79*. 96-100.

Klem, L. (2000). Structural equation modeling. In Grimm L. G. & Yarnold, P. R.. (Eds.) *Reading and Understanding More Multivariate Statistics* (pp 227-260). Washington, DC: American Psychological Association.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: The Guilford Press.

Liu, X. (2004, October). *The initial validation of teacher's perception of grading practices*. Paper presented at the 2004 Northeastern Educational Research Association annual meeting, Kerhonkson, NY.

Liu, X., O'Connell, A.A., & McCoach, D.B. (2006, April). *The initial validation of teachers' perceptions of grading practices*. Paper presented at the Annual Conference of the American Educational Research Association (AERA), San Francisco, CA.

McMillan, J. H., & Lawson, S. R. (2001). *Secondary science teachers' classroom assessment and grading practices.* Metropolitan Education Research Consortium, Richarmond, VA. (ERIC Document Reproduction Service N. Ed 450 158)

McMillan, J. H., Myran S., & Workman D. (2002). Elementary teachers' classroom assessment and grading practices. *The Journal of Educational Research, 95*, 203-213.

McMillan, J. H., & Nash, S. (2000, April). *Teacher classroom assessment and grading practice decision making*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

McMunn, N., Schenck, P., McColskey, W. (2003 April). *Standards-based assessment, grading, and reporting in classroom: Can district training and support change teacher practice?* Paper presented at the Annual Meeting of American Educational Research Association. Chicago, IL.

Stiggins, R. J., Frisbie, D. A., & Griswold, P. A. (1989). Inside high school: Building a research agenda. *Educational Measurement: Issues and Practice, 8*, 5-14.

Yuan, K.-H., & Bentler, P. M. (2007). Structural equation modeling. In C. R. Rao, & S. Sinharay (Eds.). *Handbook of Statistics 26: Psychometrics* (pp. 297- 358). Amsterdam: Elsevier Science B. V.

Table 1. Descriptive statistics of teacher's age and teaching experience by country

Country	Variable	N	Minimum	Maximum	Mean	Standard Deviation
U.S.	Age	118	21	64	35.83	11.96
	Experience	121	1	37	10.09	10.02
China	Age	153	22	66	31.84	7.08
	Experience	158	1	40	9.10	7.55

Table 2. Summary of Fit Indices from Confirmatory Factor Analysis for Full-sample Data (n=389)

Model	χ^2 (df)	p	χ^2/df	RMSEA (90% CI)	CFI
Unmodified hypothesized model	1650.50 (725)	<.001	2.277	.067 (.062, .071)	.79
Modified hypothesized model 1	1567.68 (720)	<.001	2.177	.064 (.060, .068)	.81
Modified hypothesized model 2	949.03 (473)	<.001	2.006	.059 (.054, .065)	.88

Table 3. Factors and the Corresponding Items for the Unmodified Hypothesized Model

Factor 1: Importance
1. Grading is an important criteria for judging students' progress.
2. Grading has an important role in classroom assessment.
3. Grading has a positive effect on students' academic achievement.
4. Grading practices are important measures of student learning.
5. Grading practices are important measures of student achievement.
6. Grading has a strong impact on students' learning.
Factor 2: Usefulness
7. Grading helps me categorize students as above average, average and below average.
8. Grading can help me improve instruction.
9. Grading can encourage good work by students.
10. Grading helps me in deciding what curriculum to cover.
11. Grading is a good method for helping students identify their weaknesses in a content area.
12. Grading can keep students informed about their progress.
13. Grading provides information about student achievement
14. Grading documents my instructional effectiveness
15. Grading provides feedback to my students
16. High grades can motivate students to learn
Factor 3: Student effort
17. I consider student effort when I grade.
18. I give higher report card grades for students who show greater effort.
19. I will pass a failing student if he or she puts forth effort.
20. Grades are based on students' completion of homework.
21. Grades are based on the degree to which students participate in class.
22. Grades are based on a student's improvement.
Factor 4: Student ability
23. I consider student ability in grading.

24. Grades are based on students' problem solving ability.
25. Grades are based on students' critical thinking ability.

26. Grades are based on students' independent thinking ability.
27. Grades are based on students' collaborative learning ability.
28. Grades are based on students' writing ability.

Factor 5: Teachers' grading habits

29. I tend to use letters (e.g., A, B, C) rather than numbers (e.g. 95%) in grading.
30. If a student fails a test, I will offer him/her a second chance to take the test.
31. I often give students opportunities to earn extra credit.
32. I often look at the distribution of grades for the whole class after I finish grading.
33. I have my own grading procedure.
34. I often confer with my colleagues on grading criteria.

Factor 6: Perceived self-efficacy of grading process

35. Grading is the easiest part of my role as a teacher.
36. It is easy for me to recognize strong effort by a student.
37. It is easy for me to assess student achievement with a single grade or score.
38. It is easy for me to rank order students in terms of achievement when I am grading.
39. It is difficult to measure student effort.
40. Factors other than a student's actual achievement on a test or quiz make it difficult for me to grade.

Table 4. Summary of Fit Indices from Confirmatory Factor Analysis for U.S.-sample Data (n=122)

Model	χ^2 (df)	p	χ^2/df	RMSEA (90% CI)	CFI
Unmodified hypothesized model	1104.15 (725)	<.001	1.523	.066 (.058, .073)	.79
Modified hypothesized model 1	1074.70 (720)	<.001	1.493	.064 (.056, .072)	.80
Modified hypothesized model 2	684.65 (473)	<.001	1.447	.061 (.051, .071)	.87

Table 5. Summary of Fit Indices from Confirmatory Factor Analysis for China-sample Data (n=167)

Model	χ^2 (df)	p	χ^2/df	RMSEA (90% CI)	CFI
Unmodified hypothesized model	1494.76 (725)	<.001	2.062	.080 (.074, .086)	.77
Modified hypothesized model 1	1446.92 (720)	<.001	2.010	.078 (.072, .084)	.78
Modified hypothesized model 2	893.18 (473)	<.001	2.008	.073 (.066, .080)	.85

Table 6. Fit Indices of Five Nested Models of Multi-group Confirmatory Factor Analysis (n=389)

Model	χ^2 (df)	p	χ^2/df	RMSEA (90% CI)	CFI
Unconstrained Model	1577.87 (946)	<.001	1.668	.048 (.044, .052)	.86
Measurement Weights Model	1668.25 (973)	<.001	1.715	.050 (.046, .054)	.84
Measurement Intercepts Model	2052.66 (1006)	<.001	2.04	.060 (.056, .064)	.76
Structural Covariances Model	2124.45 (1027)	<.001	2.069	.061 (.057, .065)	.75
Measurement Residuals Model	2388.02 (1067)	<.001	2.238	.066 (.062, .069)	.70

Table 7. Equivalent and Nonequivalent Factor Loadings of Items across Countries

Item	Related Factor	$\Delta\chi^2$ (df =1)	Probability
Item 1	*Importance*	1.841	> .10
Item 2	*Importance*	.481	> .25
Item 3	*Importance*	.713	> .25
Item 4	*Importance*	3.255	> .05
Item5	*Importance*	.220	> .25
Item 6	*Importance*	20.016	< .001**
Item 7	*Usefulness*	.610	> .25
Item 8	*Usefulness*	.589	> .25
Item 9	*Usefulness*	4.268	< .05*
Item 10	*Usefulness*	0	> .99
Item 11	*Usefulness*	2.25	> .25
Item 12	*Usefulness*	4.367	< .05*
Item 13	*Usefulness*	7.521	< .01**
Item 14	*Usefulness*	.159	> .25
Item 15	*Usefulness*	3.446	> .05
Item 16	*Usefulness*	3.156	> .05
Item 17	*Student effort*	1.660	> .10
Item 18	*Student effort*	.076	> .25
Item 19	*Student effort*	2.056	> .10
Item 20	*Student effort*	4.533	< .05*
Item 21	*Student effort*	.054	> .25
Item 22	*Student effort*	2.665	> .05
Item 23	*Student ability*	5.082	< .05*
Item 24	*Student ability*	.090	> .25
Item 25	*Student ability*	3.431	> .05
Item 26	*Student ability*	3.95	=.05
Item 27	*Student ability*	1.406	> .10
Item 28	*Student ability*	.653	> .25
Item 30	*Teachers' grading habits*	1.507	> .25
Item 31	*Teachers' grading habits*	1.507	> .25
Item 35	*Grading self-efficacy*	1.262	> .25
Item 37	*Grading self-efficacy*	2.273	> .10
Item 38	*Grading self-efficacy*	1.231	> .25

*Significant at p<.05; **Significant at p<.01

Table 8. Factor Loadings of Items across Countries in the Unconstrained Model

Item	Related Factor	US	China
		Regression Weights	Regression Weights
Item 1	*Importance*	1.000	1.000
Item 2	*Importance*	.758	.927
Item 3	*Importance*	.840	1.016
Item 4	*Importance*	.950	.866
Item5	*Importance*	.867	.951
Item 6	*Importance*	.060	.756
Item 7	*Usefulness*	.827	.955
Item 8	*Usefulness*	1.000	1.000
Item 9	*Usefulness*	.793	1.016
Item 10	*Usefulness*	.714	.709
Item 11	*Usefulness*	1.098	.904
Item 12	*Usefulness*	.895	.730
Item 13	*Usefulness*	.944	.672
Item 14	*Usefulness*	.930	.972
Item 15	*Usefulness*	.710	.813
Item 16	*Usefulness*	.790	1.014
Item 17	*Student effort*	1.000	1.000
Item 18	*Student effort*	2.336	1.306
Item 19	*Student effort*	3.097	1.099
Item 20	*Student effort*	.406	1.260
Item 21	*Student effort*	.644	1.309
Item 22	*Student effort*	1.327	1.223
Item 23	*Student ability*	1.000	1.000
Item 24	*Student ability*	5.828	1.037
Item 25	*Student ability*	6.566	.970
Item 26	*Student ability*	6.689	1.014
Item 27	*Student ability*	5.097	.968
Item 28	*Student ability*	4.048	.790
Item 30	*Teachers' grading habits*	1.000	1.000
Item 31	*Teachers' grading habits*	.48	1.368
Item 35	*Grading self-efficacy*	1.000	1.000
Item 37	*Grading self-efficacy*	1.137	1.432
Item 38	*Grading self-efficacy*	1.318	1.216