International Journal of Assessment Tools in Education
2022, Vol. 9, Special Issue, 300–316
https://doi.org/10.21449/ijate.1115317
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Click here for the Turkish version of this article.

Adaptation of teachers' perceptions of grading practices scale to Turkish and examination of measurement invariance

Yesim Ozer Ozkan[1,*], **Meltem Acar Guvendir**[2], **Emre Guvendir**[3]

[1]Gaziantep University, Faculty of Education, Department of Educational Sciences, Türkiye
[2]Trakya University, Faculty of Education, Department of Educational Sciences, Türkiye
[3]Trakya University, Faculty of Education, Department of Foreign Languages Education, Türkiye

ARTICLE HISTORY

Received: May 11, 2022
Revised: Oct. 19, 2022
Accepted1: Nov. 1, 2022

Keywords:
Grading practices,
Scale adaptation,
Teachers' perceptions,
Student achievement,
Measurement invariance.

Abstract: The purpose of this research is to adapt the Teacher Perceptions of Grading Practices Scale into Turkish and to examine the measurement invariance. This scale, which examines teachers' perceptions of grading methods, has six components: importance, usefulness, student effort, student ability, teachers' grading patterns, and perceived self-efficacy of the grading process. Before adapting the scale, permission was first acquired from the researcher who developed it. To ensure linguistic comparability, bilingual translators were recruited in the second phase. The semantic, experiential, conceptual, and idiomatic equivalence between the two variants of the scale were evaluated. The original and adapted scales were administered to a group of English teachers twice at a predetermined interval, and the consistency between the two applications was analyzed due to the fact that the language employed in the original test was a widely spoken group. Confirmatory Factor Analysis (CFA) was used to examine the factor structure of the original scale. Cronbach's α and McDonald's ω coefficients were calculated for the reliability of the data obtained from the scale. Finally, the measurement invariance of the scale according to gender was examined by using Multiple Group Confirmatory Factor Analysis (MGCFA), and it was determined that the measurement model fulfilled the criteria of complete gender-group invariance.

1. INTRODUCTION

Measurement and evaluation are intertwined processes that entail detection and decision-making. While measuring entails observing certain circumstances, events, or things and describing the findings with numbers or symbols, evaluation is making a decision based on an objective or criterion associated with the measurement obtained at the end of this process. In this respect, no evaluation can be made without measurement. Teachers must conduct measurements in order to make judgments about their students' achievement. With this in mind, they aim to elicit information regarding their students' achievement with the tests or assignments they have utilized.

*Corresponding Author: Meltem Acar Guvendir ✉ meltemacar@trakya.edu.tr Trakya University, Faculty of Education, Department of Educational Sciences, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

Measurement and evaluation are primarily concerned with student achievement. The purpose of post-instruction evaluation is to measure and interpret the change in student behavior induced by the teaching activities. The performance of students is compared to established guidelines or norms. As a result of the evaluation, feedback is provided for all instructional components, and quality feedback is typically the most important part of learning. (Biggs, 2001; Eraut, 2004). At this stage, what matters is that feedback is provided on time, sufficiently, and consistently. (Harlen, 2005; Serban, 2004).

Throughout this process, the teacher attempts to provide pupils with feedback regarding their progress based on the grades they have earned. Grading is the method of allocating a student to a continuum based on impressions, evidence, or a combination of the two (Anderson, 2018). But, what is the purpose of grading? Is it absolutely required to assign grades to students in order to evaluate them?

Campbell (1921) claimed that grading serves two critical functions. The first objective is to urge students to exert greater effort, and the second goal is to offer teachers information to help them improve their instruction. Bailey & McTighe (1996) stated that a third aim of grading is to provide information about student learning to a variety of populations that need and/or require information about how well students are learning or advancing in order to make appropriate judgments about them. The grades serve as a means of disseminating student success to students, parents, teachers, postsecondary institutions, and employers.

Salend and Duhaney (2002) further extended the purposes of grading to achievement, progression, effort, comparison, instructional planning, program effectiveness, motivation, communication, education and career planning, relevance, and accountability. The grading procedure serves as a demonstration of the teacher's knowledge of the program objectives. Simultaneously, the teacher can ascertain the students' learning issues and tailor their instruction to their specific needs. Thus, the program's effectiveness can be determined. Grading is used to track students' progress in learning over time, to compare students' competencies, and to monitor students' progress and efforts. This way, feedback may be provided to families with students and the level of support required can be determined. Thus in this manner, grading enables students to develop career strategies. Finally, grades are used to determine whether or not a student is eligible to graduate from a program. Consequently, indicators of academic achievement can be provided.

The teacher's role in this evaluation process is to select the behaviors that best reflect a student's progress, to develop and implement measurement methods, and to interpret the results appropriately (Küçükahmet, 2005). Gardner *et al.* (1997) identified the following critical points that a teacher should consider when assigning grades:

1. Explain the school's grading system to students in advance.
2. State explicitly the grading rules and requirements.
3. Assign grades based on objective evidence.
4. Ascertain that pupils comprehend the examination guidelines.
5. Connect the questions to what is being taught in class.
6. Never tolerate student cheating.
7. Ascertain that the exam grades are appropriate for the intended purpose.
8. Whenever possible, never alter the grade assigned.
9. Make every effort to share the exam results as soon as possible.

Furthermore, Masters (1987) and Messick (1984) emphasized the need to embrace students' evolving and partially correct ideas rather than label them as 'wrong.' According to them, it is

critical to focus on each student's individual development rather than compare them to one another.

The question of how to evaluate students fairly has long been an intriguing one, both theoretically and practically, particularly for psychologists (Meyer, 1908). A student's grade is a summary of his/her accomplishments. Notifying students of this grade level can also be handled separately. Because while a grade may motivate students to learn or boost their self-confidence, it may also have the opposite impact, diminishing the student's desire to learn or disrupting their psychology. In addition to the variables that teachers must consider when assigning grades, Gardner *et al.* (1997) proposed that the following aspects should be emphasized when notifying students of their grades.

1. If students have concerns or reservations regarding their grades, explain the reasons,
2. Inform students about the grading criteria.
3. Notify the students' parents through letter, either individually or as a group.
4. Avoid being abrasive in your provisions.
5. Maintain a balance of oral, written, and multiple-choice examinations.
6. Keep in mind that each grade should provide an opportunity for students to remedy their weaknesses.

Another thing to keep in mind is that it is important to tell the student not just her grade but also how she can improve her performance (Masters, 1987; Messick 1984).

Numerous studies have been carried out in the literature on the extent to which teachers follow the important points stated above by Gardner *et al.* (1997). These studies show that most teachers do not know how to appropriately evaluate or grade students (Brewer & deMarrais, 2015). This is especially true for teachers working in regions where the need for teachers is high and socio-economic income is low (Redding & Smith, 2016). Due to teachers' lack of training on this issue, teachers determine students' grades based on variables other than evidence of student performance (Guskey, 2015). his combination of student accomplishment and process variables can lead to score pollution that does not correctly reflect students' grades, as well as impede academic mastery and access to accurate information about academic achievement by students, families, and other education system stakeholders (Green, Johnson, Kim, & Pope, 2006).

Although teachers agree that grades should not be assigned for non-academic subjects (Frisbie, Diamond, & Ory 1979), Guskey & Bailey (2001) and Andersson (1998) argue that teachers generally avoid assigning grades solely on the basis of achievement and that when they do, they consider other factors in addition to success. Brookhart *et al.* (2016) suggest similarly that grades are typically a composite of numerous factors that teachers value (e.g., effort, ability, study habits, engagement, and participation), and that these factors vary significantly depending on what teachers believe. McMillan, Myran, & Workman (2002) used the term "chaotic grading" to refer to this type of grading. Guskey and Link (2019) propose that integrating both achievement scores and process evaluation results in end-of-term grades may result into score pollution that fails to acknowledge the information on academic competence.

A grade may represent academic achievement alone (Bailey & McTighe, 1996) or some combination of academic achievement and one or more other factors (e.g., effort, attendance, classroom participation, and/or behavior). It is much easier to interpret a grade that represents only academic achievement. If grades are based on a combination of scores from key exams, essays, quizzes, projects, and reports, as well as evidence from homework, punctuality in delivering assignments, classroom participation, study habits, and effort, the result will be a mess (Guskey, 2011).

In school, teachers decide which students pass or fail based on their grades, which are mostly determined by the written exams that students take (Koç, 1981). However, most teachers do not possess the necessary skills to assure the validity of the measurement tools they employ (Öztürk, 1988). Teachers, in particular, struggle with developing questions that are appropriate for their students' levels (Acar Güvendir & Özer Özkan, 2016). Furthermore, teachers' grades are inconsistent, regardless of whether they utilize answer keys or not when grading written examinations (Kan, 2005). Additionally, teachers might incorporate success or external factors into their measurement and evaluation processes (Semerci, 1993; Topal, 2020). According to the Ministry of National Education's [MoNE] (2005) report, the "monitoring and evaluating learning and development" competence area has the lowest average on the self-assessment scale used to evaluate teachers' self-evaluation of the qualifications included in the draft "teaching profession general competences." In other words, teachers frequently feel insecure about their measurement and evaluation abilities. Similarly, studies show that teachers in several sectors of elementary, secondary, and high school education lack measurement and evaluation skills (Adıyaman, 2005; Çakan, 2004; Erdal, 2007; Erdemir, 2007).

As a result, fair, transparent, and effective grading procedures and methods are required to aid all students in reaching higher academic standards. However, it is apparent that teachers are incompetent at all stages of the grading process, from the development of the measurement tool through its implementation. When teachers grade students, they also take into account a variety of variables other than the grade. In this context, it is important to discover teachers' perspectives on grading processes. To investigate teachers' perceptions, Liu (2004) and Liu, O'Connell, and McCoach (2006) developed the "The Teachers' Perceptions of Grading Practices Scale" in English and Chinese. The purpose of this study is to construct a Turkish version of this scale whose validity and reliability have been established in different cultures.

It is also significant to look for evidence of measurement invariance, which is required for group comparisons based on the modified "Teachers' Perceptions of Grading Practices Scale". Since the validity and reliability are based on the measurements obtained from the measurement tool, the test and item statistics calculated to obtain information about the level of validity and reliability only reflect the characteristics of the individuals in the group (Crocker & Algina, 1986). As a result, the evidence regarding the validity and reliability of measures taken in different groups may vary. The psychometric properties of the measurements acquired may be a result of the individuals' unique features or they may be a result of the measurement tool. Thus, measurement invariance investigations disclose the circumstances under which observed variables are valid and reliable between groups (Vandenberg & Lance, 2000). The other goal of this study is to find out if the measuring tool can be used to compare different groups. To do this, a measurement invariance study will be done on the "Teacher Perceptions of Grading Practices" across gender groups.

As a result, this scale, which was adapted and whose measurement invariance was investigated between groups, might be utilized as a tool in future intercultural comparisons of teachers' grading practices. This scale may also be used to make different decisions concerning the grading processes of teachers working in Türkiye. As a consequence, it was deemed necessary to investigate the scale's validity and reliability, as well as its measurement invariance.

2. METHOD

In this section, the scale adaptation steps are explained in detail. The following steps were followed for scale adaptation (Deniz, 2007; Hambleton, 1996; Hambleton, Meranda, & Spielberger, 2005; Hambleton & Patsula, 1999).

1. Permission has been received for the adaption study.
2. Field specialists were consulted on the scale's adaptability.

3. Measurement specialists were consulted on the scale's adaptability.

4. To ensure language comparability, translators fluent in both cultures were chosen. Two translators performed the translation, and the translated version of the scale was reviewed and approved by three translators.

5. A back-translation was done.

6. It was determined if the two variants of the scale were semantically, experientially, conceptually, and idiomatically equivalent.

7. A pilot application was conducted.

8. Confirmatory factor analysis (CFA) was used to examine the factor structure of the original scale.

9. Various approaches for determining reliability were utilized.

After the adaption, measurement invariance was used to determine how teachers' responses to the scale varied by gender. In the measurement invariance process, configural, metric, scalar, and strict invariance stages were followed.

2.1. Scale Adaptation Process

The measurement tool adapted in this study is the Teachers' Perceptions of Grading Practices Scale, which was developed in English and Chinese by Liu (2004) and Liu *et al.* (2006) to determine teachers' perceptions of the practices they use in the grading process. this instrument measuring teachers' perceptions of grading practices has six factors (Importance, Usefulness, Student Effort, Student Ability, Teachers' Grading Habits, Perceived Self-efficacy of Grading Process). It is 5-point Likert rating scale (1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, and 5 = strongly agree). The fit indices for the hypothesized six-factor model with 40 items were as follows: Chi-square (χ^2) = 1562.67, degree of freedom (*df*)= 687, $p< 0.001$. Confirmatory Fit Index (CFI) = .80, Root Mean Square Error of Approximation (RMSEA) = 0.067 (90% Confidence interval of 0.062 to 0.071), and $\chi^2/df = 2.277$. The reliability coefficient of the whole scale is 0.73.

Permission to adapt the scale was received by e-mail from the researchers who developed it. In the second stage, translation and back-translation processes were carried out by researchers as well as three lecturers working in English language educators who are proficient in both cultures. The researchers examined whether the two scale forms were semantically, experientially, conceptually, and idiomatically equivalent. Due to the presence of a group that spoke the original test language, the original and adapted scales were administered twice, one month apart, to a group of English teachers, and the consistency of the two applications was investigate product-moment correlation coefficient was calculated to determine the relationship between the two scales' scores (two-term, normally distributed scores). The correlation coefficient obtained was 0.86, indicating a positive, high, and significant relationship ($p<0.05$) between the two applications. CFA was performed to examine whether the factor structure of the original scale was the same in its Turkish version. Cronbach's α and McDonald's ω reliability coefficients were estimated during the scale's reliability research.

To begin with, the CFA analysis's assumptions were tested in order to verify the scale's factor structure. First of all, it was checked whether there was missing data in the data and it was observed that there was no missing data. One of its underlying assumptions is that there are no versatile extreme values. This assumption was made using Mahalanobis distances. A total of 549 teachers responded to the scale. However, 52 outliers determined by Mahalanobis distances were removed. The second assumption is that the sample size for factor analysis must be adequate. The Kayser-Mayer-Olkin (KMO) test was used to analyze this, and because the value obtained was 0.918, the sample size was large enough for factor analysis (Leech, Barrett, & Morgan, 2005). Another assumption is normality. Since CFA is a multivariate analysis, it

requires a multivariate normality assumption. This was done by using Henze-Zirkler's test, which showed that the data did not meet the assumption of multivariate normality. (hz= 1.082; p<0.05). When the variables observed in the CFA did not show normal distribution, the WLS method was preferred since the Weighted Least Squares Method (AGL-WLS Weighted Least Square Estimation) was used as the parameter estimation method (Bollen, 1989; Schermelleh-Engel, Moosbrugger, & Müller, 2003).

CFA, which was conducted to reveal how the original factor structure of the scale was in its Turkish form, was carried out using LISREL software (v. 8.71; Jöreskog & Sörbom, 2004). Cronbach's α and McDonald's ω coefficients were calculated using Jamovi software (v. 1.8; The Jamovi Project).

Multiple group confirmatory factor analysis (MGCFA) was used for the measurement invariance of the scale according to gender groups (Jöreskog & Sörbom, 1993). For measurement invariance, configural, metric, scalar, and strict invariance models were established and the difference between the CFI and RMSEA values obtained in each model from the CFI and RMSEA values obtained with the configural invariance model was taken. ΔCFI and ΔRMSEA (Chen, 2007; Cheung & Rensvold, 2002) values were used as decision criteria in the analysis of stepwise models for measurement invariance in gender groups. According to Chen (2007), in samples larger than 300, -0.010≤ ΔCFI and ΔRMSEA≤ 0.015 values are the cut-off points for the invariance decision. These values were utilized as the cut-off point for this study to ensure that measurement invariance was attained or not. For measurement invariance, "Lavaan" (http://cran.r-project.org/web/packages/lavaan/index.html) and "semTools" (http://cran.r-project.org/web/packages/semTools/index.html) available in R software packages are used. The package (http://cran.r-project.org/web/packages/MVN/index.html) was used for multivariate normality checking.

2.2. Study Group

There are 497 teachers in the study group. In terms of gender distribution, females made up 59.8% of the study group, while males made up 40.2%. In terms of school type, 28.8% work at elementary schools, 47.9% at secondary schools, and 23.3% attend work at high schools. Associate degree instructors make up 1.6% of the research group, undergraduate teachers make up 78.1%, and graduate teachers make up 20.3%. When their distribution is examined in terms of professional seniority, 6.4% have less than one year, 9.7% have 1-3 years, 9.5% have 4-5 years, 26.8% have 6-10 years, 17.1% have 11-15 years, 14.1% have 16-20 years, and 16.5% have more than 20 years of service. The data were obtained in the spring semester of the 2020-2021 academic year.

3. FINDINGS

Descriptive statistics and reliability coefficients for the six sub-factors of the scale of teacher perceptions regarding grading practices are presented in Table 1.

Table 1. *Descriptive statistics and reliability coefficients of the scale.*

	Mean	SD	Cronbach's α	McDonald's ω
Importance	3.51	0.85	0.93	0.94
Usefulness	3.55	0.69	0.91	0.92
Student Effort	3.91	0.54	0.77	0.78
Student Ability	4.04	0.58	0.92	0.93
Teachers' Grading Habits	3.72	0.53	0.67	0.70
Perceived Self-efficacy of Grading Process	2.81	0.67	0.68	0.70

The reliability values obtained for the scale's six sub-factors were found to be high. The fact that the obtained values exceed 0.70 indicates a high degree of reliability. CFA was performed to obtain evidence of the scale's factor structure. The initial CFA involved 427 participants. The scale's 40th item was insignificant. While filling out the scale, the researchers inserted the item into the first item of the relevant factor and retested 70 respondents, taking into account the high likelihood of quitting, becoming fatigued, or responding without reading the final item. As a result of CFA performed on 497 participants, item 40 (z=-1.5, p=0.13) was excluded because it was not significant (p>0.05). The 39th item on the scale has a standardized estimation value of less than 0.30, indicating that it contributes very little to the factor. As a result, this item was removed from the scale due to its low factor load. Permission was obtained from the researcher who developed the scale at the stage of removing these items. Table 2 shows the standardized regression values and the unstandardized regression coefficients for the other 38 items.

Table 2. *Factor loadings of the scale of teachers' perceptions of grading practices.*

Factor	Indicator	Estimate	SE	95% Confidence Interval		Z	p	Stand. Estimate
				Lower	Upper			
Importance	I1	0.86	0.04	0.78	0.93	22.01	<0.00	0.82
	I2	0.79	0.04	0.72	0.86	22.30	<0.00	0.83
	I3	0.72	0.04	0.65	0.79	20.22	<0.00	0.77
	I4	0.88	0.04	0.81	0.95	25.58	<0.00	0.90
	I5	0.89	0.04	0.83	0.95	26.02	<0.00	0.91
	I6	0.82	0.04	0.75	0.89	21.77	<0.00	0.81
Usefulness	I7	0.63	0.04	0.56	0.70	17.05	<0.00	0.69
	I8	0.72	0.04	0.65	0.80	19.78	<0.00	0.76
	I9	0.60	0.03	0.53	0.66	18.43	<0.00	0.73
	I10	0.70	0.04	0.62	0.79	16.24	<0.00	0.66
	I11	0.75	0.04	0.68	0.81	21.47	<0.00	0.81
	I12	0.74	0.04	0.67	0.81	20.52	<0.00	0.78
	I13	0.75	0.04	0.68	0.82	21.22	<0.00	0.80
	I14	0.70	0.04	0.62	0.78	17.14	<0.00	0.69
	I15	0.61	0.03	0.55	0.67	20.93	<0.00	0.79
	I16	0.42	0.04	0.33	0.50	9.93	<0.00	0.44
Student Effort	I17	0.54	0.03	0.48	0.59	18.56	<0.00	0.76
	I18	0.64	0.03	0.57	0.70	18.59	<0.00	0.77
	I19	0.46	0.03	0.40	0.53	14.09	<0.00	0.62
	I20	0.32	0.04	0.23	0.40	7.12	<0.00	0.34
	I21	0.43	0.03	0.36	0.49	12.40	<0.00	0.56
	I22	0.43	0.03	0.36	0.50	12.57	<0.00	0.57
Student Ability	I23	0.50	0.03	0.45	0.55	19.64	<0.00	0.76
	I24	0.56	0.02	0.51	0.60	24.56	<0.00	0.88
	I25	0.62	0.02	0.58	0.67	27.80	<0.00	0.94
	I26	0.62	0.02	0.58	0.66	28.51	<0.00	0.95
	I27	0.60	0.02	0.55	0.64	25.84	<0.00	0.90
	I28	0.45	0.04	0.38	0.53	12.21	<0.00	0.52

Table 2. *Continues.*

Teachers' Grading Habits	I29	0.30	0.05	0.20	0.40	5.77	<0.00	0.29
	I30	0.37	0.05	0.28	0.47	7.54	<0.00	0.38
	I31	0.45	0.03	0.39	0.52	13.16	<0.00	0.62
	I32	0.45	0.03	0.39	0.52	13.93	<0.00	0.65
	I33	0.51	0.04	0.43	0.59	12.46	<0.00	0.58
	I34	0.43	0.04	0.35	0.50	11.06	<0.00	0.53
Perceived Self-efficacy of Grading Process	I35	0.55	0.05	0.45	0.65	10.74	<0.00	0.54
	I36	0.33	0.04	0.16	0.31	5.87	<0.00	0.31
	I37	0.69	0.05	0.60	0.79	14.33	<0.00	0.74
	I38	0.78	0.05	0.67	0.88	14.68	<0.00	0.77

When Table 2 is seen, the standardized estimation values for factor loadings for all items vary between 0.30 and 0.89. According to Tabachnick and Fidell (2019), factor loads should be at a minimum of 0.32. Büyüköztürk (2002) categorized a load value of 0.60 or greater as high, and 0.30-0.59 as medium. As a result, all items pertaining to the factors are significant (p<0.01), and factor loads are greater than 0.30. The model's fit index values (χ^2= 1868.10, df=650, χ^2/df=2.87, RMSEA =0.06, CFI=0.97, NNFI=0.96) were significant at the 0.05 level of significance (p<0.05). When model fit indices are evaluated, χ^2/df value (2.69) is deemed acceptable by Schermelleh-Engel *et al.* (2003) and corresponds to a moderate fit, as it is less than 5, as defined by Sümer (2000). The RMSEA value shows that the fit is acceptable. NNFI and CFI values indicate a good fit of the model. Appendix 2 shows the path diagram for the six-factor model derived using DFA.

When Appendix 2 is examined, it is noticeable that the scale of 38 items with six variables was confirmed. The gender invariance of the six-factor construct was tested using multi-group CFA analyses. Multi-group confirmatory factor analysis was performed to demonstrate that the psychometric features of the scale did not remain constant across the groups to which it would be applied (Thompson, 2004). Table 3 summarizes the results of the tested invariance stages.

Table 3. *Results of measurement invariance obtained by gender.*

Stages	$\chi2$	d	CFI	GFI	RMSEA	ΔRMSEA	ΔCFI
Configural Invariance	2544.58	1300	0.89	0.97	0.06	-	-
Metric Invariance	2568.44	1332	0.89	0.97	0.06	-0.00	0.00
Scalar Invariance	2630.89	1364	0.89	0.97	0.06	-0.00	-0.00
Strict Invariance	2707.36	1400	0.88	0.96	0.06	-0.00	-0.01

In order to determine the measurement invariance between the groups at the stages in Table 3, the difference values of the fit coefficients (ΔCFI and ΔRMSEA) were given by comparing the more limited models with the configural model. In accordance with Table 3, the fit indices as a result of multi-group CFA for configural invariance show that this stage is achieved. In other words, female and male teachers use the same conceptual perspectives in responding to scale items. The fit indices as a result of multi-group CFA for metric variance and the ΔCFI and ΔRMSEA values obtained as a result of the CFI and RMSEA difference tests were interpreted. The fit indices obtained show that the model fits well with the data. To test the metric invariance, the difference between the CFI and RMSEA values obtained in the configural invariance and metric invariance stages was examined, and it was seen that ΔCFI and ΔRMSEA for metric invariance were within acceptable limits (ΔCFI ≤0.01; ΔRMSEA ≤ 0.015). This finding shows that the factor loadings of the variables included in the model do not vary depending on a person's gender.

In the scalar invariance stage, fit indices are within acceptable limits. Scalar invariance was tested by comparing the CFI and RMSEA values obtained from configural invariance to the CFI and RMSEA values obtained from scalar invariance. When the findings were analyzed, it was discovered that the measurement model for the scale of teacher perceptions on grading processes fulfilled the scalar invariance requirement ($\Delta CFI \leq 0.01$; $\Delta RMSEA \leq 0.015$). After the scalar invariance stage, the strict invariance stage was tested.

Strict invariance fit indices are within accepted limits. The difference between the CFI and RMSEA values obtained during the configural and strict invariance phases indicated that the grading practices measurement model in gender subgroups fulfilled the strict invariance stage ($\Delta CFI \leq 0.01$; $\Delta RMSEA \leq 0.015$).

4. DISCUSSION and CONCLUSION

The purpose of this research was to analyze the validity and reliability of the Turkish version of the Teachers' Perceptions of Grading Practices Scale. CFA was performed to confirm the factor structure of the original scale in its Turkish form. Cronbach's α and McDonald's ω coefficients, which measure internal consistency, were used to check for reliability. A significant t value could not be found for the scale's 40th item (Students' engagement in the course outside of the test, social events, and other activities complicates my grading procedure.). While the scale provided satisfactory fit values, it was established that the t value for the 40th item was not significant and that the error variance for this item was also rather high. As a result, item 40 was eliminated from the scale. This item is meant to assess if instructors' non-grading status hinders their work when it comes to grading students. The reason why the item does not work in the Turkish form may be due to the attitude difference between the two cultures. Some of the teachers who answered this scale think that it is normal for them to consider their students' extracurricular situations while grading. Interviews were held with the teachers regarding this item. Teachers stated that while grading, variables other than grades (such as listening to the lecture, being respectful, doing their homework regularly) also affect their grading status. They stated that they reflect these non-academic variables on their exam scores in order to motivate students, and this is the right thing to do. This article may not function in Turkish owing to the cultural differences between the two cultures. According to several teachers who responded to this scale, it is natural for them to include their students' extracurricular activities while grading. Teachers were interviewed on this subject. Teachers indicated that during grading, they take into account aspects other than grades (such as listening to the lecture, being courteous, and doing their assignments on a consistent basis). They argued that they include these non-academic characteristics in their exam results in order to stimulate pupils, which is the correct thing to do. For instance, an English teacher at a fine arts high school remarked that she considers her students' talent while grading, and the administration even encourages them to do so. This should not be suggested in foreign literature, as it would influence the validity of the scores (Guskey, 2011; Guskey & Link, 2019). While Koç (1981) asserted that teachers largely determine their students' pass-fail status based on the results of written exams, Semerci (1993), Topal (2020), Guskey & Bailey (2001), and Andersson (1998) argue that teachers can incorporate factors outside the classroom into the measurement and evaluation process. Frisbie, Diamond, and Ory (1979) argue that grades should not be assigned for non-academic areas. Otherwise, grading will be chaotic (McMillan *et al.*, 2002) and will result in score pollution (Green *et al*., 2006).

In addition, since the factor load of item 39 was 0.15 (<0.30), this item was also removed from the scale. When the English (it is difficult to measure student effort) and translated Turkish equivalents of this item are examined, it is clear that the statement is written as a factual statement rather than a perceived self-efficacy statement. Therefore, although the item is significant, it is thought that the factor load is therefore low. Since these last two items on the

original scale did not work in the Turkish form, Liu, who developed the scale, was contacted and permission was requested to remove it. After the positive response from the scale developer, these two items were removed from the scale, and confirmatory factor analysis was found appropriate to be done again. The results obtained in the repeated analysis show that the 38-item scale is consistent with the six-factor original structure and is compatible with the data. Taking into account all of the coherence values, it is possible to conclude that the theoretical framework explains the relationships between the data acquired from the Turkish form of the scale. The internal consistency coefficients of the entire scale and its sub-factors were examined to determine reliability of the data obtained from the scale. Cronbach's α and McDonald's ω internal consistency coefficients are high on the basis of the entire scale and factors. As a result, the data acquired from the scale can be said to be consistent. As a result, the means obtained from these two groups formed by gender using this scale can be compared.

The measurement invariance of the adapted scale in different groups in terms of gender was determined by examining the ΔCFI and ΔRMSEA values obtained for the models. It was concluded that the grading practices measurement model met the condition of complete invariance because it included all of the configural, metric, scalar, and strict invariance stages in gender groups. Measurement invariance of the scale across cultures was examined by Liu (2008). In Liu's study, the factor loadings of the 39th and 40th items out of 40 items in the scale were not found to be similar in the two compared samples (China and the United States). This finding shows that the answers given to items 39 and 40 differ according to cultures. In this study, these items were removed from the scale as a result of CFA, and the measurement invariance according to gender was made over 38 items and the 38-item scale provided measurement invariance.

The study was carried out with 497 teachers. The research enlisted the help of 497 instructors. The original scale's factor structure was meant to be validated in the study, and measurement invariance in different groups was evaluated. Along with these procedures, convergent and divergent validity investigations can be carried out. Furthermore, the outcomes of studies using the adapted scale are expected to increase the evidence that the scale is both valid and reliable.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Trakya University, 25.05.2022 - 05/13.

Authorship Contribution Statement

Yesim Ozer Ozkan: Methodology, Investigation, Resources, Visualization, Software, Formal Analysis, and Writing -original draft. **Meltem Acar Guvendir:** Introduction and discussion, Writing -original draft, Methodology, Investigation, Resources, and Formal Analysis. **Emre Guvendir**: Introduction and discussion, Writing, and Proofreading.

Orcid

Yesim Ozer Ozkan https://orcid.org/0000-0002-7712-658X
Meltem Acar Guvendir https://orcid.org/0000-0002-3847-0724
Emre Guvendir https://orcid.org/0000-0003-1226-9878

REFERENCES

Acar Güvendir, M., & Özer Özkan, Y. (2016). *Practicality of measurement and evaluation course in education.* Paper presented at V. Congress of Measurement and Evaluation in Education and Psychology, 1-3 September 2016, Antalya.

Adıyaman, Y. (2005). *İlköğretim 4., 6. ve 8. sınıflarında Türkçe dersine giren öğretmenlerin ölçme değerlendirme düzeyleri [The measurement and evaluation levels of teachers teach Turkish course in 4th, 6th and 8th classes in primary school]* [Unpublished Master Thesis, Kocatepe University].

Andersson, A. (1998). The dimensionality of the leaving certificate. *Scandinavian Journal of Educational Research, 42*(1), 25-40. https://doi.org/10.1080/0031383980420102

Anderson, L.W. (2018). A Critique of grading: Policies, practices, and technical matters. *Education Policy Analysis Archives, 26*(49), 1-31. http://dx.doi.org/10.14507/epaa.26.3814

Bailey, J.M., & McTighe, J. (1996). Reporting achievement at the secondary level: What and how. In T.R. Guskey (Ed.), *Communicating student learning: 1996 Yearbook of the ASCD* (pp. 119–140). ASCD.

Bollen, K.A. (1989). *Structural equations with latent variables,* Wiley.

Brewer, T.J., & deMarrais, K. (2015). *Teach for America counter-narratives: Alumni speak up and speak out.* Peter Lang Incorporated, International Academic Publishers. https://doi.org/10.3726/978-1-4539-1556-1

Biggs, J. (2001) Assessment of student learning: Where did we go wrong? *Assessment Update, 13*(6), 6-11.

Brookhart, S.M., Guskey, T.R., Bowers, A.J., McMillan, J.H., Smith, J.K., Smith, L.F., Stevens, M.T., & Welsh, M.J. (2016). A century of grading research: Meaning and value in the most common educational measure. *Review of Educational Research, 86*(4), 803-848. https://doi.org/10.3102/0034654316672069

Büyüköztürk, Ş. (2002). Faktör analizi: Temel kavramlar ve ölçek geliştirmede kullanımı. [Factor analysis: basic concepts and using to development scale]. *Educational Administration in Theory & Practice, 32*(32), 470-483. https://dergipark.org.tr/en/pub/kuey/issue/10365/126871

Campbell, A.L. (1921). Keeping the score. *School Review, 29*(7), 510-519.

Chen, F.F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464-504. https://doi.org/10.1080/10705510701301834

Cheung, G.W., & Rensvold, R.B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233-255. https://doi.org/10.1207/S15328007SEM0902_5

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston.

Çakan, M. (2004). Öğretmenlerin ölçme-değerlendirme uygulamaları ve yeterlik düzeyleri: ilk ve ortaöğretim. [Comparison of elementary and secondary school teachers in terms of their assessment practices and perceptions toward their qualification levels]. *Ankara University, Journal of Faculty of Educational Sciences, 37*(2), 99-114. https://doi.org/10.1501/Egifak_0000000101

Deniz, Z. (2007). Psikolojik ölçme aracı uyarlama. [The Adaptation of psychological scales]. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi, 40*(1), 1-16.

Eraut, M. (2004) A wider perspective on assessment, *Medical Education, 38(1),* 803-804. https://doi.org/10.1111/j.1365-2929.2004.01930.x

Erdal, H. (2007). *2005 ilköğretim matematik programı ölçme değerlendirme kısmının incelenmesi (Afyonkarahisar ili örneği). [The investigation of measurement & evaluation parts in the new elementary school mathematics curriculum (case of Afyonkarahisar)]* [Unpublished Master Thesis, Kocatepe University].

Erdemir, Z.A. (2007). *İlköğretim ikinci kademe öğretmenlerinin ölçme değerlendirme tekniklerini etkin kullanabilme yeterliklerinin araştırılması (Kahramanmaraş örneği).*

[Searching for the secondary education teachers' competence of being able to use the techniques of measurement and evaluation (example of Kahramanmaraş)] [Unpublished master thesis, Kahramanmaraş Sütçü İmam University].

Frisbie, D., Diamond, N.A., & Ory, J.C. (1979). *Assigning course grades*, Urbana, IL: University of Illinois Office of Instructional Resources. (ERIC Document Reproduction Service No. ED285496)

Gardner, W., Demirtaş, A., & Doğanay, A. (1997). Sosyal bilimler öğretimi. [Social sciences teaching] YÖK-Dünya Bankası. MEGEP.

Green, S.K., Johnson, R.L., Kim, D., & Pope, N.K. (2006). Ethics in classroom assessment practices: Issues and attitudes. *Teacher and Teacher Education, 23*(7), 999-1011. https://doi.org/10.1016/j.tate.2006.04.042

Guskey, T.R. (2011). Five obstacles to grading reform. *Educational Leadership, 69*(3), 16-21. https://uknowledge.uky.edu/edp_facpub/6

Guskey, T.R. (2015). *On your mark: Challenging the conventions of grading and reporting.* Bloomington, IN: Solution Tree Press.

Guskey, T.R., & Bailey, J.M. (2001). *Developing grading and reporting systems for student learning*, Corwin Press.

Guskey, T.R., & Link, L.J. (2019). Exploring the factors teachers consider in determining students' grades. *Assessment in Education: Principles, Policy & Practice, 26*(3), 303-320. https://doi.org/10.1080/0969594X.2018.1555515

Hambleton, R.K. (1996). *Guidelines for adapting educational and psychological test*. National Center for Education Statistics (ED).

Hambleton, R.K. & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied Testing Technology, 1*(1), 1-30.

Hambleton, R.K., Merenda, P., & Spielberger, C. (Eds.). (2005). *Adapting educational and psychological tests for cross-cultural assessment.* Lawrence S. Erlbaum Publishers.

Harlen, W. (2005) Teachers' summative practices and assessment for learning- Tensions and synergies, *The Curriculum Journal, 16*(2), 207-223. https://doi.org/10.1080/09585170500136093

Jöreskog, K.G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Scientific Software International.

Jöreskog, K., & Sörbom, D. (2004). *LISREL [Computer Software]*. Lincolnwood, IL: Scientific Software, Inc. https://doi.org/10.1002/0471667196.ess1481

Kan, A. (2005). Yazılı yoklamaların puanlanmasında puanlama cetveli ve yanıt anahtarı kullanımının (aynı) puanlayıcı güvenirliğine etkisi [The effect of using grading scale and answer key to grader's reliability]. *Eurasian Journal of Educational Research, 20(1),* 166-177.

Koç, N. (1981). *Liselerde öğrencilerin akademik başarılarının değerlendirilmesi uygulamalarının etkinliğine ilişkin bir araştırma [A research on the effectiveness of the applications of evaluating the academic achievement of students in high schools*] [Unpublished Master Thesis, Ankara University].

Küçükahmet, L. (2005). *Öğretimde planlama ve değerlendirme [Planning and evaluation in instruction].* Nobel Yayınları.

Leech, N.L., Barrett, K.C., & Morgan, G.A. (2005) *SPSS for intermediate statistics,use and interpretation* (2nd Edition). Lawrence Erlbaum.

Liu, X. (2004). *The initial validation of teacher's perception of grading practices*. Paper presented at the 2004 Northeastern Educational Research Association annual meeting, Measuring Teachers' Perceptions 14.

Liu, X. (2008). *Assessing measurement ınvariance of the teachers' perceptions of grading practices scale across cultures*. NERA Conference Proceedings 2008. 3. https://opencommons.uconn.edu/nera_2008/3.

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

Masters, G. (1987). *New views of student learning: Implications for educational measurement.* Research working paper 87.11. University of Melbourne: Centre for the Study of Higher Education.

McMillan, J.H., Myran, S., & Workman, D. (2002). Elementary teachers' classroom assessment and grading practices. *The Journal of Educational Research, 95*(4), 203–213. https://doi.org/10.1080/00220670209596593

Ministry of National Education's [MoNE] (2005). *EARGED ilköğretim 1.-5. sınıf pilot uygulama sonuçlarının değerlendirilmesi [EARGED primary education 1.-5. Evaluation of class pilot application results. MoNE Publications.*

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement, 21*, 215-237. https://doi.org/10.1111/j.1745-3984.1984.tb01030.x

Meyer, M. (1908). *The grading of students. Science*, 28(712), 243-250. https://doi.org/10.1126/science.28.712.243

Öztürk, B. (1988). *Lise sosyal bilimler dersleri öğretmenlerinin başarı testi hazırlamadaki yeterliliklerine ilişkin bir araştırma [A research on the competencies of high school social science teachers in preparing achievement tests]* [Unpublished Master Thesis, Gazi University].

Redding, C., & Smith, T.M. (2016). Easy in, easy out: Are alternatively certified teachers turning over at increased rates? *American Educational Research Journal, 53*(4), 1086-1125. https://doi.org/10.3102/0002831216653206

Salend, S.J., & Duhaney, L.M.G. (2002). Grading students in inclusive settings. *Teaching Exceptional Children, 34*(3), 8-15. https://doi.org/10.1177/004005990203400301

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8*(2), 23-74. http://www.mpr-online.de

Semerci, Ç. (1993). *Fırat Üniversitesinde öğrenci başarısının ölçülmesinde kullanılan yöntemler ile ölçme ve değerlendirmeye ilişkin görüşler [Opinions on the methods used in measuring student achievement at Fırat University and on measurement and evaluation]* [Unpublished Master Thesis, Fırat University].

Serban, A.M. (2004) Assesment of student learning outcomes at the institutional level. *New Directions For Comminity Colleges, 2004*(126), 17-27. https://doi.org/10.1002/cc.151

Sümer, N. (2000). Yapısal eşitlik modelleri: Temel kavramlar ve örnek uygulamalar [Structural equation modeling: Basic concepts and applications]. *Türk Psikoloji Yazıları, 3*(6), 49–74.

Tabachnick, B.G. & Fidell, L.S. (2019). *Using multivariate statistics* (7th edition). Pearson.

The jamovi project (2021). *Jamovi*. (Version 1.8) [Computer Software]. Retrieved from https://www.jamovi.org

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications.* American Psychological Association. https://doi.org/10.1037/10694-000

Topal, T. (2020). Öğretmen adaylarının bakış açısından sınıf öğretmenlerinin öğretim sürecinde gösterdikleri dönüt ve düzeltme davranışları [Feedback and correction behavior of the classroom teachers during the teaching process from the perspective of the teacher

candidates]. *OPUS International Journal of Society Researches, Eğitim ve Toplum Özel Sayısı, 16*, 6150-6166. https://doi.org/10.26466/opus.825157

Vandenberg, R.J., & Lance, C.E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4-70. https://doi.org/10.1177/109442810031002

APPENDIX

Appendix 1. *Teachers' perceptions of grading practices scale* (Turkish version).

Faktör 1. Önem	**Kesinlikle Katılmıyorum**	**Katılmıyorum**	**Nötr**	**Katılıyorum**	**Kesinlikle Katılıyorum**
1. Not verme, öğrencilerin gelişimlerini değerlendirmek için önemli bir ölçüttür.					
2.Not verme, sınıf içi ölçme ve değerlendirmelerde önemli bir role sahiptir.					
3. Not verme, öğrencilerin akademik başarıları üzerinde olumlu bir etkiye sahiptir.					
4.Not verme uygulamaları, sınıf içi öğrenmelerin önemli ölçülerini oluşturur.					
5.Not verme uygulamaları, öğrenci başarısının önemli ölçümleridir.					
6. Not verme, öğrencilerin öğrenmeleri üzerinde güçlü bir etkiye sahiptir.					
Faktör 2. Yarar					
7. Not verme, öğrencileri ortalamanın üstünde, ortalama düzeyde ve ortalamanın altında olarak sınıflandırmama yardımcı olur.					
8. Not verme, öğretim yöntemimi geliştirmeme yardımcı olur.					
9.Verilen notlar öğrencileri iyi çalışmalar yapmaya teşvik edebilir.					
10. Not verme, hangi konuları öğreteceğime karar vermeme yardımcı olur.					
11. Not verme, öğrencilerin bir dersin içeriğindeki zayıflıklarını belirlemeye yardımcı olan iyi bir yöntemdir.					
12.Not verme, öğrencileri gelişimleri hakkında bilgilendirebilir.					
13.Not verme, öğrenci başarısı hakkında bilgi verir.					
14.Not verme, benim etkili bir öğretim uyguladığımın bir göstergesidir.					
15.Not verme, öğrencilerime geri bildirim sağlar.					
16.Yüksek notlar, öğrencileri öğrenmeye motive edebilir.					
Faktör 3. Öğrenci Çabası					
17.Not verirken öğrencinin çabasını göz önünde bulundururum.					
18.Daha fazla çaba gösteren öğrencilere daha yüksek karne notları veriyorum.					
19.Başarısız bir öğrenciyi çaba göstermesi halinde geçiririm.					
20.Verdiğim notlar, öğrencilerin verilen ödevleri tamamlayıp tamamlamadıklarına dayanır.					
21.Verdiğim notlar, öğrencilerin sınıfta derse katılma düzeylerine dayanır.					
22. Verdiğim notlar, öğrencinin gelişim düzeyine dayanır.					
Faktör 4: Öğrenci Yeteneği					
23.Not verirken öğrencilerin yetenek düzeylerini göz önünde bulundururum.					
24.Not verirken, öğrencilerin problem çözme yeteneğini göz önünde bulundururum.					
25.Not verirken, öğrencilerin eleştirel düşünme yeteneğini göz önünde bulundururum.					

26.Not verirken, öğrencilerin bağımsız düşünme becerilerini göz önünde bulundururum.					
27.Not verirken, öğrencilerin işbirliğine dayalı öğrenme yeteneğini göz önünde bulundururum.					
28.Not verirken, öğrencilerin yazma becerilerini göz önünde bulundururum.					
Faktör 5: Öğretmenlerin not verme alışkanlıkları					
29. Not verirken, imkanım olsaydı, rakamlardan ziyade harfleri (örn., A, B, C) kullanma eğiliminde olurdum.					
30. Bir öğrenci sınavda başarısız olursa, ona sınava girmek için ikinci bir şans daha sunarım.					
31.Öğrencilere sıklıkla ek puan kazanma fırsatı veririm.					
32.Not vermeyi bitirdikten sonra sıklıkla tüm sınıfın not dağılımına bakarım.					
33.Kendime özgü not verme yöntemim var.					
34.Değerlendirme ölçütleri konusunda sık sık meslektaşlarımla görüş alışverişinde bulunurum.					
Faktör 6: Not verme sürecinin algılanan öz-yeterliği					
35.Not verme, öğretmen olarak işimin en kolay parçasıdır.					
36.Bir öğrencinin çok çaba gösterdiğini fark etmek benim için kolaydır.					
37.Öğrenci başarısını tek bir notla veya puanla değerlendirmek benim için kolaydır.					
38.Not verirken, öğrencileri başarı açısından sıralamak benim için kolaydır.					

Appendix 2. *The path diagram of factor loadings of the scale of teachers' perceptions of grading practices.*