LEARN Journal: Language Education and Acquisition Research Network
(ISSN: 2630-0672 (Print) | ISSN: 2672-9431 (Online)
Volume: 14, No: 1, January – June 2021
Language Institute, Thammasat University
https://so04.tci-thaijo.org/index.php/LEARN/index

Language Classroom Assessment Fairness: Perceptions from Students

Matthew P. Wallace[a,*], Coral Yiwei Qin[b]

[a] mpwallace@um.edu.mo, Department of English, Faculty of Arts and Humanities, University of Macau, Macao SAR
[b] Coral.Qin@uottawa.ca, Faculty of Education, University of Ottawa, Canada
[*] Corresponding author, Matthew P. Wallace, mpwallace@um.edu.mo

APA Citation:
Matthew, P. W., & Coral, Y. Q. (2021). Language classroom assessment fairness: Perceptions from students. *LEARN Journal: Language Education and Acquisition Research Network, 14*(1), 492-521.

Received 28/07/2020

Received in revised form 04/11/2020

Accepted 27/11/2020

Keywords
fairness, justice, L2 classroom assessment

Abstract

This study investigated second language learners' fairness perceptions of a classroom-based language test and whether those perceptions were predictive of justice judgments of their language program. Classroom test fairness was conceptualized as a multidimensional construct, consisting of distributive fairness (how fairly test scores represent performance), procedural fairness (how equally test procedures are applied), and interactional fairness (how respectful communication is between students and teachers during a test). In total, 192 Chinese EFL learners at a university in Macau completed an online questionnaire eliciting their fairness and justice perceptions of a single testing event. The students reported that their test administration was procedurally and interactionally fair, but were neutral about its distributive fairness. Results from structural equation modeling indicated that the students made justice judgments about their language programs based on how respectfully the teachers communicated with them during the testing event (interactional fairness) and how fairly their score

	represented their performance (distributive fairness). Perceptions of procedural fairness was not predictive of language program justice, but it was predictive of how fairly students viewed their scores. The findings suggest that language teachers should ensure that their test administrations have distributive, procedural, and interactional fairness.

1. INTRODUCTION

Ensuring that an assessment is fair is a fundamental concern for any teacher. A test may be considered fair when (1) the procedures used to administer it are the same for each student (Kunnan, 2018; McNamara & Ryan, 2011; Song, 2016; Wallace, 2018; Xi, 2010), (2) the communication between the teacher and students is respectful and clear, and (3) the score adequately represents the performance given on the test (Wallace, 2018). This conception of fairness is subjective in nature because it is determined by perceptions of the students and teachers. Fairness may also be determined objectively by examining the psychometric qualities of test scores, particularly whether or not the test was biased in favor of or against a group of test takers (Kunnan, 2018). If test scores reveal a bias, then that test administration and use of its scores would be considered unfair. Of the two perspectives of fairness, the subjective evaluation is more appropriate for classroom teachers who have limited resources to conduct advanced statistical analysis on their test data.

A closely related concept to fairness is justice and the two terms have been defined differently in the literature. In studies that have examined justice judgments of undergraduate courses and their instructors, justice has been defined as the perceived fairness of the procedures, outcomes, and interactions between learners and teachers (e.g., Chory-Assad & Paulsel, 2004b; Chory-Assad, Horan, & Houser, 2017; Horan, Chory-Assad & Goodboy, 2010). Defined this way, if the procedures, assigned grades, and communication with the teacher are perceived to be fair, then the educational practice of the teacher is considered just. In the language assessment literature, fairness and justice are distinct constructs. McNamara and Ryan (2011) propose that fairness and justice are associated with two dimensions of Messick's (1989) framework for evaluating test score validity. McNamara and Ryan consider fairness to be involved with the Evidential Basis of score use and

interpretation and justice to be involved with the Consequential Basis of score use and interpretation. To provide evidence of score validity, test data must show that (1) the test's procedures are administered equally, (2) the test measured its intended constructs, and (3) the technical properties of the test were sound. These aspects of validity can be examined objectively with the use of advanced statistical techniques. Evaluations of fairness are thus aspects of the test and its administration. In contrast, the consequences of using test scores may be determined by identifying the implicit values measured on a test and the consequences of using the test scores. Justice, therefore, is an appraisal of the institutions that administer the tests and use their scores.

Making explicit the distinction between fairness and justice, Kunnan (2000; 2004; 2018) argues that only a test and its administration may be evaluated for being fair or not and that justice judgments are reserved for the organization that delivers that assessment and/or uses its scores. Kunnan states that fair tests give learners sufficient opportunities to learn the assessed material, elicit consistent and meaningful interpretations, are unbiased, and are administered equally for all test takers. He considers just organizations to be those that utilize tests that benefit society and make the reasoning for use of a test and its scores explicit. These distinctions, though more expansive than earlier definitions of fairness and justice, still conceive the concepts as being objective in nature. This means that judgments about whether a test administration and the organization that uses its scores may be fair or just are made solely by people external to the testing process. We feel that fairness judgments should be made by those who are most affected by a test's administration—the students. Therefore, this study adopts Wallace's (2018) conception of fairness and justice being subjective in nature. The theoretical framework of fairness and justice conceptualized for this study accounts for this.

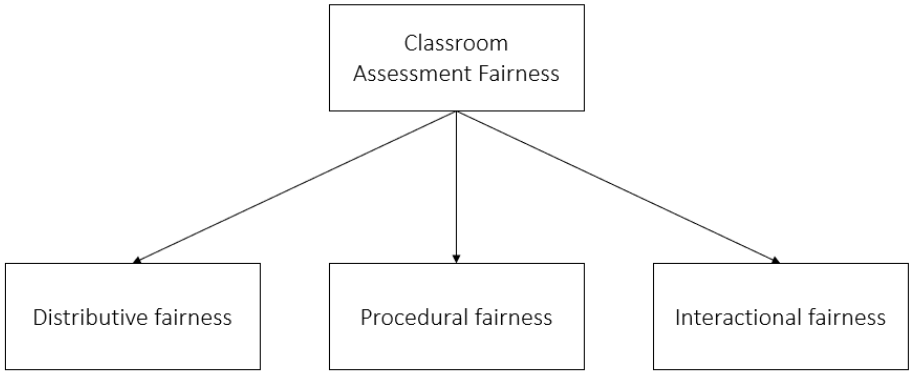
2. DIMENTIONS OF CLASSROOM ASSESSMENT FAIRNESS

Classroom language assessment fairness is a multi-dimensional construct consisting of three dimensions, namely distributive fairness, procedural fairness, and interactional fairness (see Figure 1; Wallace, 2018). Distributive fairness refers to fairness of an outcome (i.e., test score) (Deutsch, 1985). If test takers perceive their score to fairly represent

their performance, then that test administration would have high distributive fairness. Procedural fairness refers to the fairness of the procedures carried out to arrive at an outcome (Leventhal, 1980). High procedural fairness may be achieved when the test takers perceive a test administration to have been equally administered and not biased towards or against a group of test takers. Interactional fairness refers to how fairly students are treated by their teacher (Chory-Assad, Horan, & Houser, 2017), and for language testing, how fairly the test administrator (the person or institution) treats the test taker (Wallace, 2018). If test takers perceive their interactions with their test administrator to be respectful, then that test administration would have high interactional fairness.

Figure 1

Multidimensional Language Assessment Fairness Model



The empirical literature suggests that perceptions of classroom fairness can influence student behavior. When learners perceive the grading procedures to be fair, they have been shown to be more motivated to learn and view their course instructor more favorably (Chory-Assad, 2002). Students act aggressively toward their instructor when they think the grade they received is an unfair representation of their performance (Chory-Assad, 2002; Chory-Assad, Horan, & Houser, 2017; Chory-Assad &Paulsel, 2004b) and when their interactions with their teacher are disrespectful (Chory-Assad, 2007). Most of this research has been devoted to examining the relationship between fairness

perceptions and student affect in content-area university classrooms. With the exception of Wallace (2018), the relationship between fairness perceptions of a language test administration and its perceptions of the social entity administering that test has received little attention. This is an interesting oversight considering the potential negative consequences facing teachers should they and the tests they administer be perceived as being unfair.

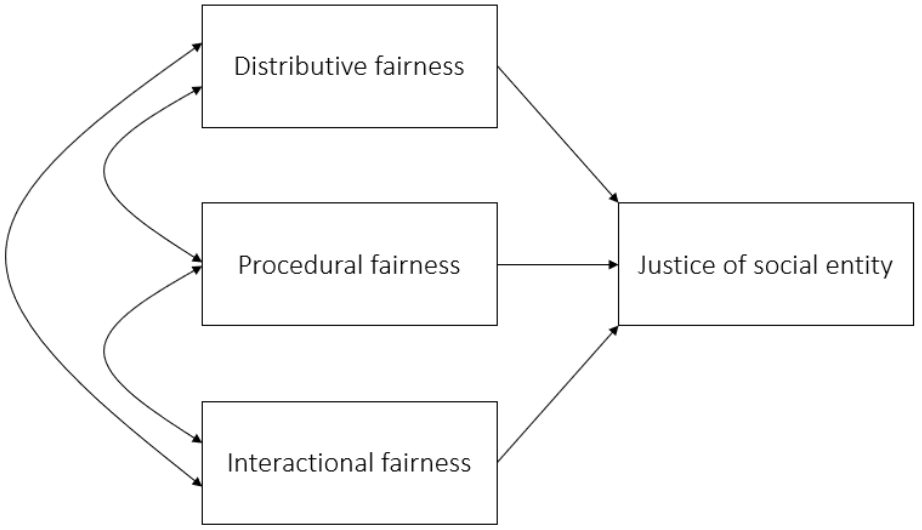
3. WALLACE (2018)

Wallace (2018) used the multi-dimensional language assessment fairness model to examine the fairness perceptions of a language test administration for second language learners in Taiwan universities. Survey data was collected from 83 university students studying varied second languages (English, Chinese, Japanese, Korean, German, French, Spanish, Indonesian, and Turkish). The results showed that students perceived their language test administrations to have high interactional fairness and high procedural fairness, but a moderate amount of distributive fairness. This means that when students completed tests in their language classes, they considered their interactions with their teacher to be respectful and the assessment procedures to have been carried out equally for all learners. However, they were neutral in their perceptions of the test score being representative of their performance. Wallace also examined which of the three dimensions of classroom assessment fairness may predict justice judgments of the language program (see Figure 2). His results showed that perceptions of procedural fairness were the strongest predictor of justice judgments of the language program. Interactional fairness was also predictive of justice judgment, though not as strong as procedural fairness. Interestingly, distributive fairness did not predict perceptions of language program justice. Wallace's results were consistent with assertions made by Chory-Assad (2002) and Chory-Assad and Paulsel (2004b) that distributive fairness may not be as important for justice as procedural fairness. As long as the procedures were carried out in an equal manner, the students in Chory-Assad (2002) and Chory-Assad and Paulsel (2004b) studies felt that their courses were fair. Wallace used his results to emphasize that the perceived fairness of one test administration can influence how justly the overall language program may be viewed. He

advised that when administering a classroom test, language teachers should ensure that the test procedures are the same for every student. When communicating with students during a test, teachers should make sure their communication is clear and respectful.

Figure 2

Model of Language Assessment Fairness and Justice Judgment of Social Entity



4. CRITICAL DISCUSSION OF WALLACE'S (2018) METHODOLOGY

There are a number of limitations in Wallace's (2018) study. First, the fairness and justice perceptions were of different tests and administrations for different languages in different institutions. We question whether examining perspectives from such varied learning contexts would have provided a valid representation of the perceptions. It is legitimately possible that perceptions of an elementary-level German grammar test with 10 multiple-choice items could have been combined with perceptions of a high-level Chinese mid-term exam with 100 items measuring their reading, writing, listening, and speaking skills of varied response types and an intermediate-level Indonesian speaking interview test that spanned five minutes. The characteristics of these tests, as well as the learning contexts in which they were administered, were not

accounted for in the study. When examining perceptions of fairness and justice, it is essential for these variables to be controlled in order for valid conclusions to be drawn from the results. This study controlled these variables and by doing so, allowed for more valid conclusions to be drawn.

When reporting the results of the analysis, the fit statistics for the confirmatory factor analysis were not provided. It is important to provide these statistics so that the reader may independently verify the quality of model fit for the data. Not providing fit statistics can raise questions regarding the construct validity of the instrument used to measure the fairness dimensions. A final limitation of the study was the use of regression to examine the relationships among the variables. Regression analysis was appropriate given the sample size of the study, but in order to rigorously examine the relationships among the variables, latent variable methods, like structural equation modeling are ideal because they eliminate error variance present in the observed variables, thus allowing for cleaner observations of the relationships among the variables.

5. THE CURRENT STUDY

This study addressed the limitations of Wallace (2018) with the aim of generalizing the results to the broader language learner population. Firstly, data were collected from a larger sample in a different context from the earlier study. Having more data points increased the statistical power of the results above that of the previous study, which strengthens the conclusions drawn in the current study. Second, this study used a revised survey based on Wallace's (2018) results. Wallace found that items measuring a fourth dimension of fairness, informational fairness, performed poorly, with a low reliability estimate (α = .487) and poor fit to a confirmatory model. Because of this weak performance, the informational fairness items were removed from the questionnaire, leaving 18 items total on the survey. Perhaps more importantly, the current study introduced a common language test upon which the participants could reflect (see Instruments below). Instead of collecting data on fairness perceptions of different language tests of different languages that were completed at different training institutions, this study investigated the perceptions of an English listening and reading test

offered by one school and was administered at the same time. The current study aimed to answer the following three research questions:

1. How do university Chinese EFL learners in Macau perceive the dimensions of classroom fairness (distributive, procedural, interactional) of a classroom language test administration?

2. Are the dimensions of classroom fairness positively related to one another? Are they positively related to justice perceptions of a language program?

3. Do the dimensions of classroom fairness predict justice perceptions of a language program?

6. METHOD

6.1 Context and participants

This study was situated in an English language center at an English medium university in Macao SAR. The aim of the English language center is to improve undergraduate students' language proficiency to be able to successfully complete undergraduate content courses. To do so, the English courses focus on the development of all four language skills in general, academic, and professional English. Students at the center are assigned to one of three levels of English courses (low, intermediate, advanced) based on their performance on a placement test delivered at the beginning of first year and course grades for subsequent years. The participants for this study were recruited from the low and intermediate level courses because these students completed a common reading and listening test at the end of their semester. This test served as the assessment used for reflection by the participants.

A total of 245 Chinese EFL learners completed the questionnaire. After cleaning the data, which consisted of removing univariate and multivariate outliers, participants who failed to honestly complete the questionnaire (straight-lined answers), and participants who elected to withdraw from the study, 192 respondents were included in the analysis. Overall, 43% (*n*=83) of the participants were male students and the rest of the participants (57%, *n*=109) were female. The number of the participants who enrolled in the intermediate-level course was 60 (31%) and the remaining 132 participants were from the low-level course

(69%). The native languages of the students were mostly Cantonese (19%, *n* = 36) or Mandarin (77%, *n* = 148).

6.2 Instruments

6.2.1 Classroom Fairness Survey

A survey questionnaire was adapted from Wallace (2018) to fit the current study in order to measure learners' perceptions of classroom fairness in three dimensions—distributive, procedural, and interactional fairness—as well as the general perceptions of justice of the language program administering the test. The questionnaire consisted of items eliciting demographic information and the test scores of the participants and 18 statements representing the four intended constructs (see Appendix A). Participants indicated their agreement with the statements using a five-point Likert scale (ranging from 1 = strongly disagree to 5 = strongly agree). Four items measuring distributive fairness elicited participants' perceptions of whether their scores on the classroom-based reading and listening test represented their performance on the test (items 1, 5, 10, and 15). Five procedural fairness items elicited perceptions of the test procedures and their equality for all test takers (items 2, 4, 6, 13, and 17). Four interactional fairness items measured perceptions of the respectfulness of communications with the persons delivering the test (i.e., their classroom instructor; items 3, 7, 11, and 14). Finally, five entity justice items gauged respondent perceptions of the overall justice of the language program (social entity) within which the test was delivered (items 8, 9, 12, 16, and 18). The bi-lingual (Chinese and English) survey was developed and distributed to the participants through Qualtrics, which is a web-based survey service that is available for use by all students in the university. Qualtrics supports PC log-in or mobile log-in which was convenient for the participants and likely to collect more responses than through email. The order of the items was varied to avoid a possible order effect. Four items were negatively worded (items 8, 9, 10, and 14) to ensure participants responded to each item honestly and avoided them from marking the same value without reading the statements. Before the administration of the questionnaires, a pilot was conducted with 20 students (10 from each level) to check the clarity of instructions, estimate the time needed to complete the survey, and allow for revisions to be made to the questionnaire. Students were

required to finish the survey questionnaire within one week after they finished the reading and listening test (described below).

6.2.2 Reading and Listening Test

The reading and listening test was a one-hour paper-pencil test designed to measure the improvement in reading and listening ability. It comprised three sections, listening section one (five items), reading section (10 items), and listening section two (five items). Altogether, the test consisted of 10 reading items and 10 listening items with a total score of 20 (one for each item). The response format for all 20 items was multiple-choice with five answer options. The listening texts were adapted from a TED talk and the reading texts were adapted from USNEWS, an online periodical, and adjusted to no more than 400 words. The Flesch Kincaid Grade Level of the texts was 8.7 which was expected to be level appropriate for the participants. Students were made aware of the date and time of the test as it was listed in the course syllabus and explained by the instructor during the first day of class. This test format was consistent with the textbook practices that the instructor used during class each week.

The test was administrated by the instructor during class time on the test day. Students were arranged in rows facing the front of the class. Answer sheets and reading texts were distributed to the students before the test began. For the listening sections, the instructor played the audio from the front of the room and the students indicated the correct answers of the corresponding questions on the answer sheet. For the reading section, students read the texts from a separate paper and then answered the corresponding questions on the answer sheet. When the test ended, answer sheets were collected and graded by the instructor. One point was awarded for each correct response. Total scores out of 20 were given to the participants on the same day of the test. In the following class meeting, the instructor distributed the answer sheet, listening transcripts, and reading texts back to the students and explained the answers to each item. The participants were made aware that they had the right to file an appeal if they identified inaccurate test score.

6.3 Analysis

The four items that were negatively worded were reverse coded so they could be on the same scale as the remainder of the items. The data then underwent a thorough screening process. First, the data was inspected for straight-lined responses, where the participants answered the same value for every item. When found, these participants were removed from the dataset. Unexpected responses were then identified using Winsteps (Linacre, 2016), a Rasch Modeling software package. Winsteps generates a probabilistic model of expected responses of each participant for each item based on the pattern of responses on an instrument. It then compares the expected pattern of responses with the actual pattern given by the participants. When the responses are unexpected, they are flagged by the program. For example, if responses to four items measuring distributive fairness were 5-5-5-1, then Winsteps would flag '1' for being unexpected because the four responses measuring the same construct should be similar. Once identified, the inconsistent responses were removed from the analysis.

The data was then checked for univariate and multivariate outliers. To identify univariate outliers, Z-scores were calculated for each item on the questionnaire and the reading and listening test score. If the Z-score was outside of the absolute value of 3.29 for a single data point, then it was considered an outlier and removed from the dataset (Tabachnick & Fidell, 2013). Multivariate normality was checked by calculating the Mahalanobis distance of the items of the questionnaire and then comparing the Mahalanobis distances against a chi-square distribution with the same degrees of freedom (Tabachnick & Fidell, 2013). If the probability that the Mahalanobis distance was different from the chi-square distribution was below .001, then that value would be considered an outlier and was removed from the dataset. After confirming that the data was univariate and multivariate normal, Cronbach's alpha reliability estimates were calculated to examine internal consistency of the items on the questionnaire.

To verify the factor structure of the classroom language assessment fairness questionnaire, confirmatory factor analysis was conducted. Two first-order confirmatory models were specified: a Fairness Dimensions Model and an Entity Justice Model. In the Fairness Dimensions Model, the four items measuring distributive fairness were regressed onto a Distributive Fairness factor. The five items measuring

procedural fairness were regressed onto a Procedural Fairness factor. Finally, the four items measuring interactional fairness were regressed onto an Interactional Fairness factor. For the Entity Justice Model, the five items measuring justice judgments of the language program were regressed onto an Entity Justice factor. The specified confirmatory models are presented in Appendices B and C. Fit statistics recommended by Kline (2016) were inspected to determine how well the data fit the model. Good fit is demonstrated when chi-square value is non-significant (above .05), the comparative fit index is near 1.0, and the root mean square error of approximation value is below .05. The factor loadings of each observed variable were expected to be above .600, indicating that the variables measured a similar underlying construct.

Descriptive statistics of the raw data were calculated to answer research question 1 regarding students' perceptions toward the dimensions of classroom fairness. The average score of the items under each fairness dimension and entity justice was calculated to arrive at a mean score for each variable. To answer research questions 2 and 3, structural equation modeling was conducted. The two confirmatory models from confirmatory factor analysis were combined into a structural model. The Entity Justice factor was regressed onto the Distributive Fairness, Procedural Fairness, and Interactional Fairness factors (see Appendix D). The same fit statistics recommended by Kline (2016) to evaluate model fit for the confirmatory models were used for the structural model. To answer research question 2 about the relationship among the fairness variables and entity justice, the correlations among the factors were observed. To answer research question 3 about the relative contribution of each fairness dimension to justice judgements of the language program, the standardized coefficients were examined.

7. RESULT

The results of the descriptive statistics and reliability analysis are presented in Table 1. The standard item alphas of the scales were all above .80, indicating high internal consistency. Results from confirmatory factor analysis presented in Table 2 show that the Fairness Dimensions

Model and the Entity Justice Model fit the data moderately well. The comparative fit index was near 1.0 for both models, but the chi-squared statistic was statistically significant and the root mean square error of approximation was above .05 for both models. The results presented in Appendix E also show that the factor loadings for the observed variables were above .70. These results suggest that the items designed to measure distributive fairness, procedural fairness, interactional fairness, and entity justice measured their respective latent variable.

Table 1

Descriptive Statistics and Cronbach' Alpha Reliability Estimates

	M (max 5)	*SD*	Cronbach's alpha
Distributive fairness	3.65	0.94	.88
Item 1	3.90	0.91	
Item 5	3.72	0.95	
Item 10	3.16	1.02	
Item 15	3.84	0.89	
Procedural fairness	4.12	0.85	.87
Item 2	3.90	0.80	
Item 4	4.36	0.77	
Item 6	4.18	0.86	
Item 13	4.08	0.93	
Item 17	4.08	0.91	
Interactional fairness	4.07	0.93	.90
Item 3	4.27	0.82	
Item 7	4.30	0.80	
Item 11	4.13	0.92	
Item 14	3.57	1.19	
Entity Justice	3.63	0.99	.92
Item 8	2.84	1.09	
Item 9	3.22	1.07	
Item 12	4.01	0.94	
Item 16	3.98	0.95	
Item 18	4.07	0.91	

Table 2

Model Fit Indices for First-order Confirmatory Models

Model:	X^2	*df*	*p*-value	*CFI*	*RMSEA*
Fairness Dimensions Model	153.79	62	.000	.939	.088
Entity Justice Model	76.99	5	.000	.884	.275

In response to research question 1, the descriptive statistics show that the students perceived procedural fairness to be highest (M = 4.12, SD = 0.85). Interactional fairness (M = 4.07, SD = 0.93) was also perceived to be fair and distributive fairness (M = 3.65, SD = 0.94) was the lowest reported. The students' reported overall feeling toward their language programs were neither just nor unjust (M = 3.63, SD = 0.99).

The second research question asked if the three dimensions of classroom assessment fairness are associated with one another and entity justice. The intercorrelations among the observed variables presented in Appendix F show that almost all of the observed variables are positively correlated with one another. This is consistent with the standardized correlation coefficients of the latent variables presented in Table 3 showing that the fairness dimensions shared a positive relationship with each another. The Procedural Fairness variable moderately correlated with the Distributive Fairness (r = .663, p < .001), Interactional Fairness (r = .668, p < .001), and Entity Justice (r = .660, p < .001) variables. The Distributive Fairness variable weakly to moderately correlated with the Interactional Fairness variable (r = .414, p < .001) and moderately correlated with the Entity Justice variable (r = .621, p < .001). Finally, the Interactional Fairness variable moderately correlated with the Entity Justice variable (r = .646, p < .001)

Table 3

Standardized Correlation Coefficients for the Latent Variables in the Study (N =192)

	1	2	3	4
1. Distributive	1.00			
2. Procedural	.663***	1.00		
3. Interactional	.414***	.668***	1.00	

4. Entity Justice	.621***	.660***	.646***	1.00

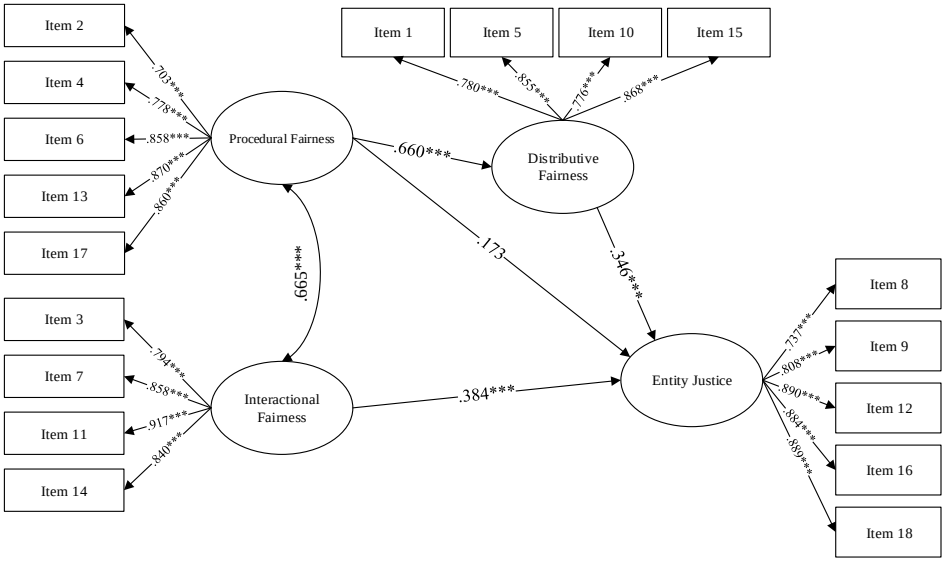
Note: *** $p < .001$

The third research question asked which dimensions of classroom assessment fairness predict justice judgments of the language program. Results from structural equation modeling show that the data fit the model moderately well: $\chi 2$ (129) = 423.94, p = .000, CFI = .877, RMSEA = .109. The results show that the three dimensions of fairness explained 60% of the variance in justice judgments of the language program, though only two dimensions were predictive. The Interactional Fairness variable (β = .388, $p < .001$) and Distributive Fairness variable (β = .347, $p < .001$) weakly to moderately predicted the Entity Justice variable. Surprisingly, the Procedural Fairness variable did not predict the Entity Justice variable, though it had the strongest bivariate correlation among the three fairness variables. This indicates that the presence of either the Distributive Fairness variable or Interactional Fairness variable in the model fully mediated the effect of the Procedural Fairness variable on the Entity Justice variable. Of the two possibilities, the Distributive Fairness variable was considered the most likely because students may perceive their score to be a fair representation of their performance if they perceive the test procedures to be unbiased. To examine this relationship, a mediation model was specified. The Distributive Fairness variable was regressed onto the Procedural Fairness variable. A bidirectional parameter was set between the Procedural Fairness and Interactional Fairness variables. The Entity Justice variable was regressed onto all three fairness variables. The specified mediated structural model for classroom assessment fairness dimensions and entity justice is presented in Appendix G.

The results show that the model fit was similar to the previous model: χ^2 (130) = 424.23, p = .000, CFI = .877, RMSEA = .109. The Distributive Fairness (β = .346, $p < .001$) and Interactional Fairness (β = .384, $p < .001$) variables had direct effects on the Entity Justice variable, and the Procedural Fairness variable had an indirect effect on the Entity Justice variable. The Procedural Fairness variable predicted the Distributive Fairness variable (β = .660, $p < .001$) and positively moderately correlated with the Interactional Fairness variable (r = .665, $p < .001$). The mediated structural model of classroom assessment fairness dimensions and entity justice presented in Figure 3.

Figure 3

Mediated Structural Model of Classroom Assessment Fairness Dimensions and Entity Justice with Standardized Estimates of Strength of Relationships



Note. Ovals represent latent variables and rectangles represent observed variable item groups. Direction of arrows indicates directionality of effect.
***$p < .001$

8. DISCUSSION

The students reported that the reading and listening test administration was procedurally and interactionally fair, but held neutral stances for distributive fairness and entity justice. This means that they felt that the procedures used to administer the test were equal for every student and that the communication with their teacher about the test was respectful. However, they also thought that their test score neither fairly nor unfairly represented their performance and that their language program was neither just nor unjust. There are three possible explanations for the neutral ratings of the distributive dimension. Firstly, there was a discrepancy between the expected grade and the grade they received and the grade of their peers (Chory-Assad & Paulsel, 2004b; Chory-Assad, 2007) for many students. Chory-Assad (2007) reported that distributive fairness is mostly predicted by the grade students expect to receive.

When that grade differs from their expectation, they perceive that assessment to be less fair. Another explanation may be due to student perceptions of their teacher's character, or whether or not they are trustworthy (Chory-Assad, 2007). Chory-Assad (2007) reported that distributive fairness was predicted by perceptions of instructor trustworthiness. If the students viewed their teachers to be untrustworthy, then they would view their scores to be unfair. Of the two potential reasons for neutral distributive fairness, the expected grade is the more likely of the two. Given that the reading and listening test used in this study contained only multiple-choice items, the test itself did not require high communication frequency between the test administrator and the students before, during, or after the test. Therefore, the instructor's credibility may not have strongly influenced the students' perceptions of the distributive fairness. A third reason for neutral distributive fairness may be that the reading and listening test did not adequately measure performance. Song (2016) reported that Chinese English as a foreign language viewed fair tests to be those that are based on merit. In other words, students were more likely to perceive distributive fairness to be high if they were evaluated through performance-based assessment, as the judgments of students' knowledge and skills would be based on observation of their behavior or inspection of their production (Stiggins, 1995). However, multiple-choice questions used on the reading and listening test of this study was restrictive in terms of production. It is possible that the students felt this test format was too rigid to demonstrate their language abilities, and therefore not adequately reflect their language capability.

The results closely coincide with Wallace (2018), who reported that language learners in Taiwan viewed their language test procedures and interactions with their language teachers administering their test to be fair, but were neutral in their views of distributive fairness and entity justice. The findings of the current study and Wallace's study suggest that language teachers at university language centers may be doing a good job of treating their students respectfully when a test is being administered and that the procedures of the test are applied equally for all test takers, but they are not doing as well at showing how the test score is based on merit. This may be concerning for the language programs because the tests they deliver may be seen as unfair. The negative consequences associated with administering unfair tests include

students losing motivation to learn, viewing the teacher as less credible, and acting aggressively toward the course instructor (Chory-Assad, 2002). In order to avoid such unfavorable outcomes, teachers should do their best to ensure tests are administered fairly.

The results from correlation analysis in response to research question 2 showed that the dimensions shared a positive moderate relationship with one another and with the Entity Justice variable. These results align with Wallace (2018), who reported positive moderate relationships among each of the same variables in his study. These findings suggest that when students perceived their language test to be procedurally fair and administered equally for everyone, they also viewed their scores to be fairly representative of their performance, their interactions with the teacher delivering the test to be respectful, and the language program that delivered the test to be just. Wallace proposed that these results would be welcomed by classroom language teachers who can avoid negative perceptions of assessment fairness by ensuring their test administrations are procedurally unbiased, score distributions are based on merit, and student interactions are clear and respectful. Not doing so could lead to negative consequences (i.e., lower motivation and teacher credibility and higher aggression towards the teacher; Chory-Assad, 2002) and certainly harm the learning process and affect the outcomes of assessment.

The weakest relationship in both Wallace and the current study was between the Distributive Fairness variable and the Interactional Fairness variable. This indicates that perceiving a score to fairly represent performance may not be closely related to perceiving the communication with the teacher delivering the test to be respectful. It is possible that the nature of the test reflected on may have affected this relationship. The reading and listening test that was reflected on in this study involved limited interactions with the teacher, not much beyond a reading of the instructions and handing out and collecting test materials. If the test involved some interaction with the teacher giving the test, like an interview speaking test, then this relationship may have been stronger. Ultimately, though, the correlation results tentatively support the theoretical framework utilized in this study (see Figure 1), that the three dimensions make up different components of classroom assessment fairness.

When the variables were examined for directional relationships to answer research question 3, the results from the mediated structural model showed that the Interactional Fairness variable was the strongest predictor of the Entity Justice variable. This means that when the students appraised their language program, they considered how their teacher communicated with them during the testing event to have the strongest influence. If their teacher was respectful in their interactions, then the students viewed the language program to be just. This result partially coincided with Wallace's (2018) findings, who showed that the Interactional Fairness variable was the second strongest predictor of the Entity Justice variable. The results of both of these studies suggest that language teachers should take special care in communicating with their students during a test administration. This is especially important given that students have been shown to evaluate their teachers' credibility based on how respectful they communicate in class (Chory-Assad & Paulsel, 2004a). Chory-Assad and Paulsel (2004a) reported that teachers are viewed with high interactional fairness when they treat their students with kindness, consider the students' perspectives, and demonstrate empathy to students' feelings. This sensitivity to how the teacher communicates is understandably heightened during a testing event, when student anxiety rises.

Interestingly, the results also showed that the Distributive Fairness variable mediated the relationship between the Procedural Fairness and Entity Justice variables. This means that perceptions of how fairly the procedures of a single test was carried out did not have a direct effect on how justly students perceived their language program. Instead, the perceptions of the procedures directly affected how students viewed the credibility of their score given their performance. If students viewed their test administration to have been procedurally unbiased, then they also viewed their scores of that test to be fair representations of their effort. And if the students viewed their test scores to be fair, then they also viewed the language program to be just. These results differ from Wallace's (2018) findings, who reported that the Procedural Fairness variable was the strongest predictor of the Entity Justice variable and that the Interactional Fairness variable did so to a lesser degree. Surprisingly, the Distributive Fairness variable did not predict the Entity Justice variable in his study. Both studies reported that judgments of the language program were influenced by the communication among the

test administrators and students. This consistency suggests that Chinese learners of other languages consider respectful communication to be important when evaluating how justly language programs behave. The biggest difference between the studies' results is in how the fairness perceptions of one test administration's procedures can affect the justice judgments of the entire language program. For the language learners in Taiwan, fair test procedures were reportedly the most important consideration when judging the language program. In contrast, Macau EFL learners viewed fair test procedures to be indicative of fair score representation and did not directly affect how justly programs are perceived. In this sense, procedural fairness was not viewed as an equal dimension of classroom fairness as distributive fairness or interactional fairness, but that procedural fairness was a contributor to distributive fairness. This contradicts the theoretical framework proposed in this study (see Figure 1) suggesting that all three dimensions of classroom fairness equally contribute to justice judgments. Rather, students took into account the procedures of a test administration in order to determine if their scores were fair or not, as opposed to making a direct judgment about the language program overall based on procedural fairness perceptions. The contradictory results between the current study and Wallace's may be due to methodological limitations of the earlier study. Wallace elicited perceptions of different test administrations of different stakes in different learning contexts throughout Taiwan, which prevented a consistent reporting of fairness and justice perceptions. Without a common test administration or language program to reflect upon, the results of the survey may be somewhat misleading. Overcoming this limitation, the participants in this study reflected on the same language test and language program that was conducted near the end of the semester with the score of the test counting 20% towards participants' final grades. The relatively high stakes of the test may explain why the Distributive Fairness variable was one of the predictors of the Entity Justice variable. Students care very much about their performance on these assessments because not doing so may result in their having to continue to take courses offered by the language center. This, undoubtedly would heighten awareness to the credibility of the scores awarded for their test performance.

9. LIMITIONS AND CONCLUSION

This study makes a significant theoretical contribution to the language teaching, learning, and assessment literature. Classroom assessment fairness was conceptualized as a multidimensional construct, with each dimension being equal and distinct predictors of entity justice (see Figure 2). However, the results of this study suggest that only two dimensions had direct effects on justice judgements—distributive fairness and interactional fairness. The third dimension, procedural fairness, had an indirect effect on justice judgments through distributive fairness. This means that justice judgments of the language program were made mostly from how respectfully the teachers communicated with their students during the testing event and how fairly their score represented their performance. The appraisal of the scores were affected by how equally the test procedures were carried out as it was administered.

This study has three limitations. First, due to practical constraints, the instrument (the reading and listening test) in this study was not validated. Though this was a reflection of an authentic test in a classroom setting, a validated instrument could provide stronger empirical evidence with less irrelevant data variation. Secondly, this study was only situated in a single language program from a university in Macau. The specific context may have restricted how generalizable the results of the study may be to the greater language learner population. Thirdly, the study did not account for other possible variables in the classrooms that may have contributed to the results (e.g., teacher credibility, students' grade expected, students' learning motivation and attitude). Future studies are therefore encouraged to expand the scope of investigation to include these variables. Finally, the use of a closed-ended questionnaire restricts, to some degree, how much students are able to express themselves in regard to classroom assessment fairness. Future studies are also encouraged to utilize qualitative data-collection methods to elicit these perceptions in the students' own words.

The findings of this study have important pedagogical implications. The results indicate that teachers should take special care to ensure that test procedures are applied equally for every student so that scores can be perceived as fair. To maintain credibility for themselves and that of the language program, teachers should maintain respectful and courteous communication with students during a testing event. This advice seems intuitive, but since the possible consequences of test administrations being perceived unfairly are so great (e.g. lower

motivation, acting aggressively toward instructor, less learning), it is worth reinforcing these points.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous LEARN Journal reviewers for their suggestions for improvement. The authors would also like to thank the LASeR lab at the University of Macau for their support of the study.

ABOUT THE AUTHORS

Matthew P. Wallace: an Assistant Professor in the Faculty of Arts and Humanities in the Department of English at the University of Macau. He is the co-director of the LASeR lab and his research interests include second language listening comprehension, language assessment fairness, and learner motivation.

Coral Yiwei Qin: a Ph.D. student in the Faculty of Education at the University of Ottawa. Her research interests include language teaching and assessment, test impact, development and validation of language assessment and language assessment policy.

REFERENCES

Chory-Assad, R. M. (2007). Enhancing student perceptions of fairness: The relationship between instructor credibility and classroom justice. *Communication Education*, *56*(1), 89-105. https://doi.org/10.1080/03634520600994300

Chory-Assad, R. M., Horan, S. M., & Houser, M. L. (2017). Justice in higher education classroom: Students' perceptions of unfairness and responses to instructors. *Innovative Higher Education*, *42*(4), 321-336. https://doi.org/10.1007/s10755-017-9388-9

Chory-Assad, R. M. (2002). Classroom justice: Perceptions of fairness as a predictor of student motivation, learning, and aggression. *Communication Quarterly*, *50*(1), 58-77. https://doi.org/10.1080/01463370209385646

Chory-Assad, R. M., & Paulsel, M. L. (2004a) Antisocial classroom communication: Instructor influence and interactional justice as predictors of student aggression. *Communication Quarterly*, *52*(2), 98-114. https://doi.org/10.1080/01463370409370184

Chory-Assad, R. M., & Paulsel, M. L. (2004b). Classroom justice: Student aggression and resistance as reactions to perceived unfairness. *Communication Education*, *53*(3), 253-273. https://doi.org/10.1080/0363452042000265189

Deutsch, M. (1985). *Distributive justice: A social-psychological perspective*. Yale University Press.

Horan, S. M., Chory-Assad, R. M., & Goodboy, A. K. (2010). Understanding students' classroom justice experiences and responses. *Communication Education*, *59*(4), 453-474. https://doi.org/10.1080/03634523.2010.487282

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (Vol. 9, pp. 1-14). Cambridge University Press.

Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. Weir (Eds.), *European Year of Languages Conference Papers, Barcelona* (pp. 27-48). Cambridge University Press.

Kunnan, A. J. (2018). *Evaluating language assessments*. Routledge.

Leventhal, G. S. (1980). What should be done with equity theory? In K. J. Gergen, M. S. Greenberg, & R. H. Willis (Eds.), *Social exchange: Advances in theory and research* (pp. 27-55). Springer US.

Linacre, J. M. (2016). *Winsteps® Rasch measurement computer program*. Winsteps.com.

McNamara, T., & Ryan, K. (2011). Fairness versus justice in language testing: The place of English literacy in the Australian Citizenship Test. *Language Assessment Quarterly*, *8*(2), 161-178. https://doi.org/10.1080/15434303.2011.565438

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Education Measurement* (3rd ed., pp. 13-103). American Council on Education & Macmillan.

Song, X. (2016). Fairness in educational assessment in China: Historical practices and contemporary challenges. In S. Scott, D. E. Scott, &

C. F. Webber (Eds.), *Assessment in education: Implications for leadership* (pp. 67-89). Springer.

Stiggins, R. J. (1995). *Sound performance assessments in the guidance context*. ERIC. http://www.ericdigests.org/1996-3/sound.htm

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Pearson.

Wallace, M. P. (2018). Fairness and justice in L2 classroom assessment: Perceptions from test takers. *Journal of Asia TEFL*, *15*(4), 900-1238. http://www.doi.org/10.18823/asiatefl.2018.15.4.11.1051

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, *27*(2), 147-170. https://doi.org/10.1177/0265532209349465

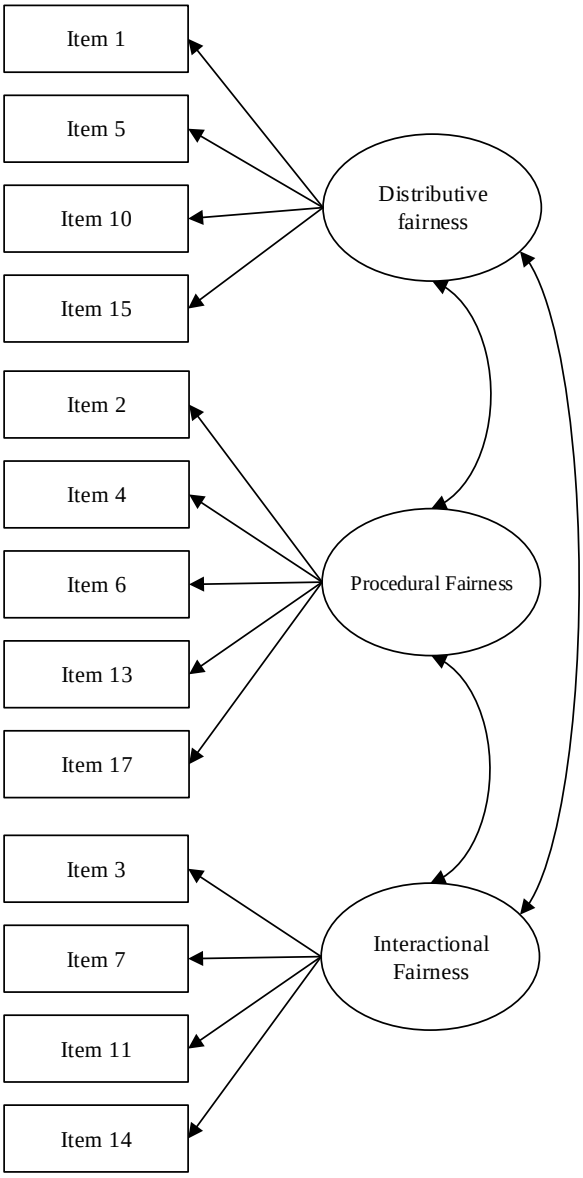
Appendix A

Classroom Fairness and Justice in L2 Assessment Questionnaire Items

1. My score reflects the effort I have put into studying for the test.
2. I have been able to express my views about the test procedures (i.e., whether I think they are fair) if I wanted to.
3. The instructor giving me the test treated me in a polite manner before the test.
4. I am able to appeal my score if I want to.
5. My score accurately reflects my performance on the test.
6. The test procedures have been applied consistently (same for all test takers).
7. The instructor giving me the test treated me with respect during the test administration.
8. The way the language program assesses students is not fair.
9. I think most of the people taking the test would say they are unfairly treated by the language program.
10. Given my performance on the test, my score is not justified.
11. The instructor treated me with dignity when I received my test score.
12. For the most part, the language program treats test takers fairly.
13. The test procedures (steps to complete the test) were fair to all test takers.
14. I have received disparaging remarks from the instructor about my test performance.
15. My score is appropriate for the performance I gave on the test.
16. In general, I can count on the language program to be fair.
17. The test procedures were free of bias (i.e., equal for all test takers).
18. Overall, the language program has treated me fairly.

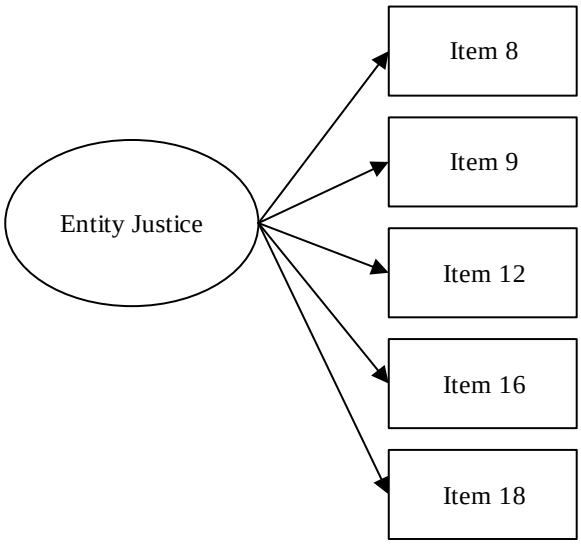
Appendix B

Specified first-order confirmatory Fairness Dimensions Model



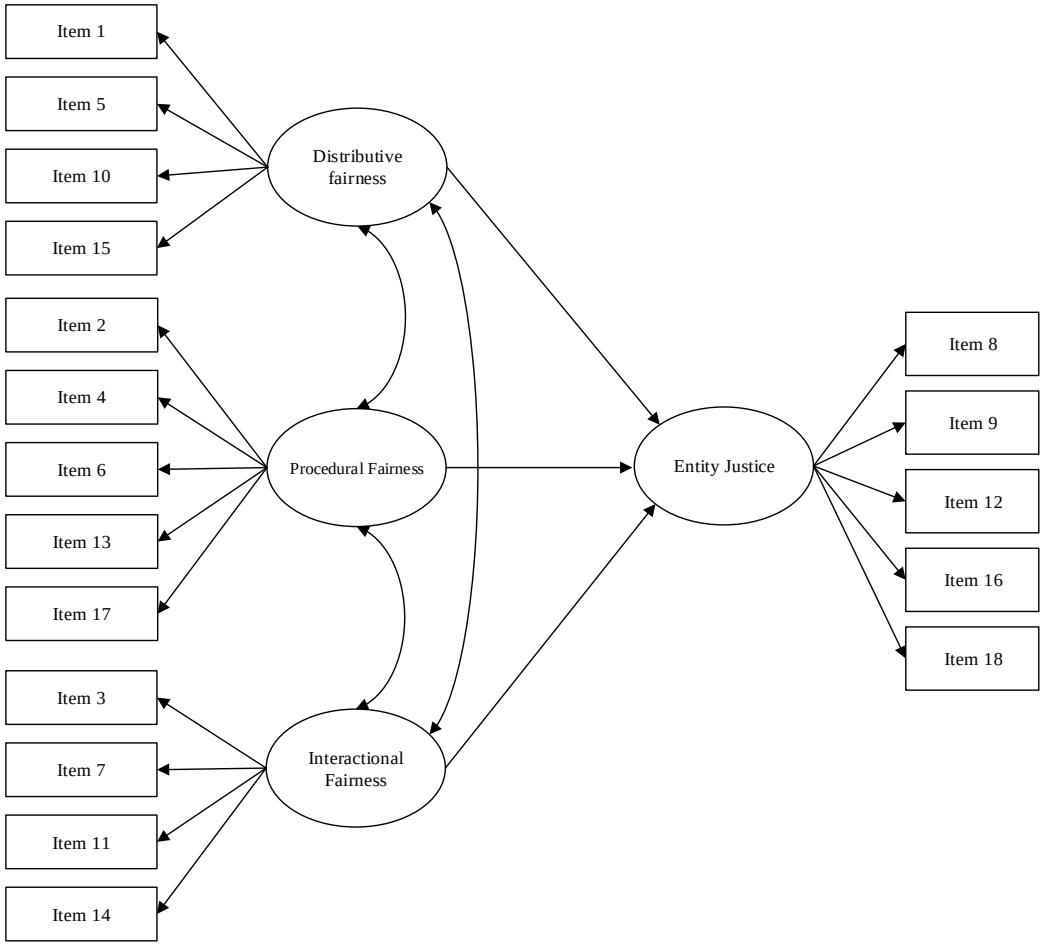
Appendix C

Specified first-order confirmatory Entity Justice Model



Appendix D

Specified structural model for classroom assessment fairness and entity justice



Appendix E

Unstandardized and standardized factor loadings for study's confirmatory factor analyses (*N* = 192)

Observed variable	Latent factor	Unstandardized (*SE*)	Standardized
Item 1	Distributive fairness	0.53 (.05)	.78
Item 5	Distributive fairness	0.60 (.05)	.86
Item 10	Distributive fairness	0.59 (.06)	.78
Item 15	Distributive fairness	0.59 (.05)	.87
Item 2	Procedural fairness	0.59 (.06)	.70
Item 4	Procedural fairness	0.62 (.05)	.78
Item 6	Procedural fairness	0.75 (.05)	.86
Item 13	Procedural fairness	0.81 (.06)	.87
Item 17	Procedural fairness	0.78 (.06)	.86
Item 3	Interactional fairness	0.65 (.05)	.79
Item 7	Interactional fairness	0.72 (.05)	.86
Item 11	Interactional fairness	0.83 (.05)	.92
Item 14	Interactional fairness	1.00 (.08)	.84
Item 8	Entity justice	1.00 (.00)	.74
Item 9	Entity justice	1.07 (.11)	.81
Item 12	Entity justice	1.03 (.10)	.89
Item 16	Entity justice	1.07 (.10)	.88
Item 18	Entity justice	0.99 (.09)	.89

Note. All loadings were significant at $p < .001$;

Appendix F

Standardized correlation coefficients for the observed variables

Standardized correlation coefficients for the observed variables

Measure	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. Item 1	1.00																
2. Item 5	.72***	1.00															
3. Item 10	.61***	.64***	1.00														
4. Item 15	.63***	.71***	.67***	1.00													
5. Item 2	.36***	.39***	.35***	.51***	1.00												
6. Item 4	.31***	.40***	.43***	.48***	.65***	1.00											
7. Item 6	.41***	.46***	.41***	.48***	.61***	.68***	1.00										
8. Item 13	.42***	.55***	.45***	.58***	.51***	.55***	.75***	1.00									
9. Item 17	.35***	.45***	.42***	.59***	.60***	.63***	.70***	.79***	1.00								
10. Item 3	.25***	.25***	.34***	.38***	.39***	.48***	.48***	.39***	.35***	1.00							
11. Item 7	.19**	.27***	.38***	.40***	.40***	.59***	.48***	.45***	.45***	.71***	1.00						
12. Item 11	.17*	.27***	.31***	.48***	.50***	.54***	.55***	.52***	.52***	.69***	.80***	1.00					
13. Item 14	.13	.20**	.37***	.32***	.44***	.44***	.47***	.42***	.43***	.73***	.71***	.75***	1.00				
14. Item 8	.23*	.31***	.58***	.40***	.30***	.34***	.30***	.39***	.44***	.38***	.35***	.39***	.46***	1.00			
15. Item 9	.28***	.33***	.62***	.50***	.28***	.37***	.28***	.39***	.39***	.43***	.41***	.47***	.47***	.85***	1.00		
16. Item 12	.38***	.50***	.42***	.58***	.51***	.57***	.51***	.66***	.63***	.49***	.55***	.63***	.50***	.61***	.65***	1.00	
17. Item 16	.37***	.45***	.40***	.53***	.41***	.39***	.43***	.59***	.54***	.44***	.44***	.55***	.49***	.59***	.67***	.80***	1.00
18. Item 18	.31***	.39***	.39***	.59***	.41***	.42***	.43***	.56***	.60***	.37***	.43***	.53***	.41***	.63***	.68***	.78***	.82***

Notes. Items numbers refer to items on the questionnaire. Sections divided into distributive fairness items (1-4), procedural fairness items (5-9), interactional fairness items (10-13), and entity justice items (14-18). *$p < .05$; **$p < .01$; ***$p < .001$

Appendix G

Specified mediated structural model for classroom assessment fairness dimensions and entity justice