Journal of Psychoeducational Assessment

Assessing measurement invariance of the Student Personal Perception of Classroom Climate (SPPCC) across different ethnic groups

Journal:	*Journal of Psychoeducational Assessment*
Manuscript ID	JPA-14-0024.R5
Manuscript Type:	Regular Article
Keywords:	Ethnicity < Culture/crosscultural, Structural equation modeling < Measurement, Elementary school < Participants, School climate < Social and educational environment, Elementary education/childhood < Social and educational environment

Abstract: The class climate is acknowledged as being related to student learning. Students learn more in classrooms that are supportive and caring. However, there are few class climate instruments at the elementary school level. The aim of the current study was to assess the measurement invariance of a recently-developed scale in a different context (New Zealand) from where it was developed (US) and across different ethnic groups. A total of 1924 elementary school students (963 males and 961 females) participated. Students completed the Student Personal Perception of Classroom Climate (SPPCC). Results of the multiple-group CFA contrasting the SPPCC with four ethnic samples (New Zealand European, Māori, Pasifika, and Asian) indicated that the SPPCC represented the same four factors in classroom climate (CC) for each of these groups (configural invariance). Results also revealed that full metric invariance was supported although only partial scalar invariance was achieved because of a lack of invariance in the thresholds for five items. Therefore, this study provided empirical support for the SPPCC when used within a new context and with different ethnic groups. Future studies to enhance the usability of the SPPCC are discussed.

Assessing measurement invariance of the Student Personal Perception of Classroom Climate (SPPCC) across different ethnic groups

Abstract

The class climate is acknowledged as being related to student learning. Students learn more in classrooms that are supportive and caring. However, there are few class climate instruments at the elementary school level. The aim of the current study was to assess the measurement invariance of a recently-developed scale in a different context (New Zealand) from where it was developed (US) and across different ethnic groups. A total of 1924 elementary school students (963 males and 961 females) participated. Students completed the Student Personal Perception of Classroom Climate (SPPCC). Results of the invariance tests of the SPPCC across four ethnic samples (New Zealand European, Māori, Pasifika, and Asian) indicated that the SPPCC represented the same four factors in classroom climate (CC) for each of these groups (configural invariance). Results also revealed that full metric invariance was supported although only partial scalar invariance was achieved because of a lack of invariance in the thresholds for five items. Therefore, this study provided empirical support for the SPPCC when used within a new context and with different ethnic groups. Future studies to enhance the usability of the SPPCC are discussed.

Keywords: class climate, measurement, ethnic groups

Assessing measurement invariance of the Student Personal Perception of Classroom Climate (SPPCC) across different ethnic groups

Introduction

The classroom climate is a "global summary of the psycho/social/emotional and organizational/managerial state of the classroom" (Babad, 2009, p.54). It encompasses the learning environment created through the pedagogical beliefs and instructional activities of the teacher, and the management and organization of that environment. Classroom climate also relates to teacher–student and student–student relationships. Often, classroom relationships result from the way they are framed by the teacher. For example, when students change seating groups regularly, it is more likely they will form relationships across the classroom. When students sit in ability groupings, it is more likely they will form friendships based on those groupings (Author, 2015). Thus, although teachers may not be solely responsible for all classroom relationships, they certainly contribute to them (Babad, 2009). Similarly, how learning is structured contributes to the classroom climate. When achievement is made salient and students are encouraged to perform at higher levels than their peers, the classroom climate is likely to be different from classes where students are working towards improving individual skills and collaboration is encouraged (Anderman, Patrick, Hruda, & Linnenbrink, 2002).

Classroom climate measurement arose out of earlier work by Moos (1979) and Walberg (1979). Fraser (1986) applied this earlier work to secondary school science classrooms and led the early classroom climate research within secondary contexts. Even today, there are fewer instruments to measure classroom climate in elementary schools than there are measures for secondary school classrooms. The current study tested whether empirical support could be found for the recently developed classroom climate scale (Rowe,

Kim, Baker, Kamphaus, & Horne, 2010), designed for elementary school classrooms, when employed in a new context from that in which it was developed, and with different ethnic groups.

Defining Classroom Climate

Definitions of the classroom climate vary as do the dimensions used to measure the classroom climate. Generally, the classroom climate is considered to be the psychosocial environment in which learning occurs (Rowe et al., 2010). The classroom climate has been measured using observations, or low inference measurement, whereby trained observers have recorded students' and teachers' interactional behaviours (Peterson & Walberg, 1979). Researchers (e.g., Fraser, Anderson, & Walberg, 1982) have gathered students' perceptions of the overall classroom climate using self-report measures that include several factors (high inference measurement). There are also researchers who have assessed the classroom climate through qualitative methods such as interviews and observations (e.g., Weinstein, 2002). Most commonly, however, self-report measures have dominated the literature (Babad, 2009). These self-report measures use several dimensions depending on the scale, but together the dimensions indicate a student's perceptions of the overall class climate. Student interpretations are considered a useful measure of the classroom climate because what students perceive is likely to affect their beliefs about, and reaction to, the classroom climate.

Classroom climate has been associated with various student outcomes and, therefore, has been considered an important classroom dimension. Indeed, while schools focus on measuring student academic outcomes, psychosocial outcomes have been considered even more significant by some (Babad, 2009; Fraser, 1986). For example, the overall assessment of the classroom climate has been associated with student motivation. In reviewing the relations between motivation and the classroom climate, Urdan and Schoenfelder (2006) focused on three theoretical perspectives of motivation: achievement goal theory, self-

determination theory, and social-cognitive theory. They argued that the emphasis teachers place on achievement as performance, rather than achievement as mastering skills, leads to teachers making the success criteria salient to students. In turn, students are likely to adopt performance or mastery orientations depending on the criteria they perceive the teacher emphasizes. Self-determination theory focuses on the needs of students: need for competence, autonomy, and relatedness. Deci and Ryan (2002) have shown that students' needs can be fulfilled, enhanced, or thwarted depending on their perceptions of the overall classroom climate. Self-efficacy theory (Bandura, 1993) highlights student success as influencing beliefs about the probability of achieving particular skills, and of future success. The teacher contributes meaningfully to student self-efficacy through providing messages about whether or not students are likely to achieve new learning. Student perceptions are influenced by the achievement of their peers, reinforcement from teachers, information gained vicariously, and characteristics of teachers (Urdan & Schoenfelder, 2006).

Student engagement has also been linked with overall perceptions of the classroom climate. Furrer and Skinner (2003) reported that student perceptions of their relatedness to teachers and peers predicted their overall perceptions of the classroom climate. When views of the classroom climate changed, there were corresponding changes in student engagement, particularly emotional engagement. Further, students' perceptions of the classroom climate declined from fifth to sixth grade, and this affected student engagement. Adolescent identity formation and behaviour are also influenced by the classroom climate. Roeser, Eccles, and Sameroff (2000) reported that early adolescents decide to engage in learning based on whether they feel competent to complete their classroom tasks, whether they value the activities and understand their purpose, and whether they feel cared for by teachers – one dimension of all class climate measures. Therefore, teachers were viewed as being able to enhance student motivation, by ensuring that students understood the value and purposes of

assignments, and through offering students emotional support and encouragement, particularly when students found an academic task difficult.

The overall classroom climate has also been associated with the development of student social skills. Brophy-Herb, Lee, Nievar, and Stollak (2007) showed that teachers who rated students' social competence negatively or positively were shown to display either negative teacher behaviours and poor classroom support, or positive class behaviours and evaluations of teacher behaviours. A positive classroom climate has been associated with the development of student social competence. Hamre and Pianta (2001) tracked students from kindergarten to eighth grade and examined the importance of the teacher–student relationship for future student outcomes. They found that, accounting for demographic variables and initial behaviour ratings and ability, students whose relationship was rated as poor by teachers at kindergarten had lower achievement and more behavioural difficulties at eighth grade.

One reason there has been increasing interest in the classroom climate is because it has been associated with student achievement. Students from ethnic minority groups are often those whose relationships with teachers are poor (McKown & Weinstein, 2008). McCormick, O'Connor, Cappella, and McClowry (2013) explored achievement outcomes for Black and Hispanic students when they had close relationships with their kindergarten teachers, and found that high-quality relationships in kindergarten predicted mathematics achievement in grade one. Studies, such as these, indicate that there may be differences in the ways that different ethnic groups view the class climate.

However, psychosocial constructs such as class climate have frequently been applied to all groups, without regard for culture or ethnicity. Psychologists have employed measures across contexts, assuming that they can be universally applied, and will provide similar findings when used in a different context (Bullen & Rubie-Davies, 2013). There have been several studies in the motivation field (e.g., Hornstra, van der Veen, Peetsma, & Volman,

2013) that have shown differences by ethnicity in student reports of their motivation. In one study examining the goal orientation of students from different ethnic groups, Freeman, Gutman, and Midgley (2002) found that among older elementary school students, African American students espoused both mastery and extrinsic goals more than White students. African American students were motivated to learn and master skills, but also wanted to achieve good grades. However, African American students seldom endorsed performance goals, despite being otherwise extrinsically motivated. Further, when they viewed their classrooms as supporting extrinsic goals, their self-efficacy was higher and they showed greater self-regulated learning than White students.

From the literature, class climate measures do not appear to have been examined in relation to different ethnic groups. The class climate scale used in the current study, for example, has not previously been investigated in relation to responses from different ethnic groups, even though there is evidence to suggest that variations may be found. Nunn (2011), for example, has shown that whereas some teachers encourage cooperation and friendships across racial lines, others do not and, instead, subconsciously or otherwise, promote racial division. Students' perceptions of student academic and personal support in these two types of classrooms are likely to vary among ethnic groups, and particularly among ethnic minority groups. Similarly, in classes in which there were racial divisions, Nunn (2011) reported that teachers interacted more frequently with, and were more friendly towards White, rather than other students. Again, it would appear likely that in these contrasting classrooms, ethnic minority students may perceive differences in student academic and personal support.

Variations may also be found among different ethnic groups of students in terms of their perceived competence, a further dimension included in the SPPCC. Schweinle and Mims (2009) investigated student mathematics self-efficacy in classrooms that were predominantly African American or White. They found that there was no difference in

African American students' self-efficacy dependent on context. However, among White students, self-efficacy was much higher when they were in largely African American classrooms than when they were in majority White classrooms. Further, in situations such as those described by Nunn (2011), in which minority groups were most often found in lower tracks in high school, several of the students reported that Latino students were less competent than White students. Because ethnic minority students are often relegated to the lower tracks, not only in the US but also in New Zealand (Author, 2015), it is possible that they may have decreased perceptions of their competence compared with those of White students.

The final dimension measured in SPPCC is student satisfaction with school. Again, it would seem possible that in schools where ethnic minority groups do not feel welcome (see Nunn, 2011), that they would be less satisfied with school. Indeed, Ennis (1998) reported that in classes where African American students believed their teachers respected them and cared about their learning, the students were far more satisfied with their schooling than they were in classes where they believed teachers made little effort to prepare lessons, and were often dismissive of the students. There have been similar findings in New Zealand, whereby Bishop and Berryman (2006) have shown that Māori students often believed that teachers did not respect their culture and had low expectations of them, leading to students expressing dissatisfaction with their schooling.

Measuring Classroom Climate

Thus, the strong association of the classroom climate with many other classroom variables highlights the importance of the classroom climate for student academic and psychosocial development. However, most measures have been designed for the secondary school context, for example, the Learning Environment Inventory (Fraser et al., 1982), the Classroom Environment Scale (Moos & Trickett, 1987), and the Constructivist Learning

Environment Survey (Taylor, Fraser, & Fisher, 1997). Fraser et al. (1982) developed the My Class Inventory (MCI) to be used with students aged 8 to 12 years but until recently this has been the only measure suitable for young students. Furthermore, many items in the MCI are phrased at the whole class level, for example, "In our class the work is hard to do". Using secondary school data, Fraser, Giddings, and McRobbie (1995) showed that students' perspectives of their personal experiences were different from their reports of their classroom experiences. Due to their cognitive development, younger students have more difficulty making holistic judgments than older students. That is, younger students are much more able to report their personal experience of a situation than they are able to step back and evaluate perceptions at the whole class level (Weinstein, 2002). Thus, there is a need for an instrument which can measure student personal perceptions of the classroom climate, because this is more likely to reflect students' actual perceptions of the classroom climate than if they are asked to take a more distal, whole class perspective.

Rowe et al. (2010) recently produced the Student Personal Perception of Classroom Climate (SPPCC) scale. This instrument asks for individual perceptions of the classroom climate rather than a global summary, and is aimed at younger students. Because student perceptions of the classroom climate at the classroom level are likely to be a less accurate reflection of their actual perceptions than a direct measure of student personal perceptions of their classroom experiences, for the reasons outlined above, the SPPCC measures student perceptions at the individual level. Further, given that the SPPCC is currently the only class climate measure available for elementary school students that measures their personal perceptions, it is a very useful addition to the measurement field. The SPPCC is based on Fraser's MCI (Fraser et al., 1982), which has had much empirical measurement support, but also accounts for the recent classroom climate literature, providing an instrument based both in a well-used scale, but also incorporating recent research that now informs the class climate

field. Considering the limited number of classroom climate measures for younger students, further empirical support for the SPPCC would seem opportune. Further, to our knowledge the SPPCC has not been examined in relation to differing ethnicities. Examining instruments, such as the SPPCC, with different ethnic groups in Western countries is useful because minority groups are more vulnerable to the classroom climate than majority students, and often experience a more negative classroom climate than their peers (McKown & Weinstein, 2008). Testing the perceptions of minority and majority groups, provides teacher feedback in relation to student perceived biases. Also, it cannot be assumed that all groups will have similar perceptions of the classroom climate. The current study aimed to examine classroom climate perceptions across four different ethnic groups, using the SPPCC. The main research question was: To what extent is the SPPCC a useful measure of the classroom climate in a multi-ethnic community?

The New Zealand Context

New Zealand provides an ideal context for the testing of the validity of the SPPCC across several ethnic groups because it has four distinct groups, each of which makes up a sizeable proportion of the total. The current study was conducted in Auckland, New Zealand's largest city, where among the adult population, 59% are New Zealand European, 11% are Māori, 15% are Pasifika and 23% are classified as Asian (Statistics New Zealand, 2013). Among the school-age population, however, the proportion of the smaller groups is greater (Statistics New Zealand, 2010). Thus, Auckland has a growing multicultural community.

Māori are the indigenous population. Pasifika students are those who identify their heritage as one of the Pacific Islands (e.g., Samoa, Tonga, Fiji, Cook Islands), although often the students are second or third generation born in New Zealand. Asian students are defined as those originating from South East Asia, China, and India. Most of these students are

recent immigrants. New Zealand European students are mostly white and primarily of British origin.

In New Zealand, compulsory schooling consists of two levels: primary and secondary. The current study concerns students in the primary (elementary in the US) setting. Students generally attend primary school from Years 1 to 6 (ages 5 to 10 years). At primary school, students mostly have one teacher who teaches all academic subjects. Years 7 and 8 are considered to be within the primary realm. Some primary schools called 'full primary' have Years 1 to 8 in one school. However, many Year 7 and 8 students attend what are termed 'intermediate schools', which cater exclusively for these two year levels. At Years 7 and 8, although students will still have one teacher for all core curriculum areas, they may have specialist teachers for subjects such as music, art, physical education, and technology.

Class climate research in New Zealand is nascent. There have been very few studies that have measured the class climate quantitatively (see Author, 2015 for one exception) but a few that have written about class climate based on other measures, observations, or qualitative findings (e.g., Author & colleague, 2011; Bishop & Berryman, 2006). Information pertaining to the class climate is important for elementary school teachers in the New Zealand context because both Māori and Pasifika students underachieve, and many years of pedagogical interventions have done little to narrow this gap (Hattie, 2008). It may be that psychosocial explanations for the achievement gap in New Zealand may prove more fruitful than pedagogical explanations. This could lead to intervention studies aimed at improving the class climate for New Zealand students, and particularly for Māori and Pasifika students. The current study enabled testing of the construct validity of SPPCC. Should it be shown that students from all groups indeed perceive the scales similarly, and yet one or more groups have more negative perceptions, this would provide evidence on which future interventions could be based.

Materials and Method

Participants

The participants were 1924 students, aged 7 to 12, enrolled at 12 New Zealand primary (elementary) schools. Of the schools, 3 were in low and high socioeconomic areas, respectively, and 6 were in middle-income areas. The sample was 49.9% female (n = 961). Representation by class level was respectively from Years 3 to 8 (approximately 7-12 years): 5.7%, 18.5%, 18.5%, 17.7%, 19.2% and 20.4%. Of the students, 905 (47%) were New Zealand European, 362 (18.8%) were Māori (the indigenous group), 313 (16.3%) were Pasifika (those from the Pacific Islands), 284 (14.8%) were Asian (those from South East Asia and the Indian subcontinent), and 60 (3.1%) were from other ethnic groups. The sample size for the "other" ethnicity group was not large enough to conduct Structural Equation Modelling analyses, and so was excluded from further analyses.

Procedure

Having gained permission for student participation, researchers administered the SPPCC in classrooms. Class teachers were not present so that student responses were confidential to the researchers. In each class, a researcher read out the items so that any students with reading difficulties would not be disadvantaged. At the same time, research assistants were available to provide support for students having difficulty. Two research assistants were used with younger students (Years 3-5) and one in classes of older students (Years 6-8). These procedures ensured a high level of student completion.

Instrument

The Student Personal Perception of Classroom Climate (SPPCC; Rowe et al., 2010) measures the classroom climate among elementary school students. The SPPCC has 26 items

belonging to six scales: Teacher Academic Support, Teacher Personal Support, Student Academic Support, Student Personal Support, Academic Competence, and Satisfaction. The theoretical background for this scale was based on existing models and scale items were adapted from these studies (Harter, 1985; Huebner, 1994; Johnson et al., 1983).

Rowe et al. (2010) tested a four-factor model and a six-factor model for the scale. Although the six-factor model yielded better goodness of fit and the chi-square difference test supported this, the authors selected the four-factor model because it was more parsimonious. Therefore, with support from the empirical and literature evidence, the authors combined the Teacher Academic-Teacher Personal and Student Academic-Student Personal Support scales into single dimensions: Teacher Support and Peer Support respectively. The authors also argued that elementary school students may not fully conceive the difference between academic and personal support.

The four factors of the SPPCC are measured on a four-point Likert scale as follows: 0 = *never*, 1 = *sometimes*, 2 = *often* and 3 = *always*. The internal consistency values ranged from .79 to .91, and factor correlations were between .27 and .60 (Rowe et al. 2010).

In the current study, however, students were asked to respond on a 5-point Likert-type scale: 1 = *false*, 2 = *mostly false*, 3 = *sometimes false, sometimes true*, 4 = *mostly true*, 5 = *true*. The SPPCC was part of a larger questionnaire that students completed which used a 1-5 scale. The researchers believed young students may become confused if the scale changed mid-way through the questionnaire and, therefore, all item responses were on the 1-5 scale. Further, Marsh (1990) has argued that it is important to include negative items in questionnaires so that students cannot simply check the same response for every item. Negatively worded items provide a check of the veracity of student response. Therefore, in addition to item 24 which was the only negatively worded item in the scale, four items (3, 8, 13, and 17) were also turned from a positive to a negatively worded item.

Data Analysis

Researchers use questionnaires to compare different cultural, ethnic, and gender groups' attitudes, values, or beliefs, assuming that the instrument measures the same trait in all groups. However, this tenet cannot be assumed; instead it should be tested empirically. The establishment of measurement invariance (MI) is a prerequisite for unbiased comparisons and interpretations across groups. Measurement invariance is said to exist if the observed score differences between groups only depends on the construct being measured, regardless of group membership.

We conducted the analyses in three stages. First, we checked the data for univariate outliers and missing cases. Second, the factorial structure of the SPPCC scale was examined. Finally, measurement invariance (or measurement equivalence) was tested across ethnicity, to examine whether the SPPCC constructs and items were understood and interpreted identically by different ethnic groups. Zumbo (2003) demonstrated that item-level measurement non-invariance may not manifest itself in the scale-level analyses. Therefore, we employed both scale-level and item-level invariance tests in this study.

Measurement Invariance (MI) provides evidence that, for example, the factors among SPPCC are measuring the same construct in the same way (or in a similar manner) across different ethnic groups. If the scale were to be measuring the construct differently for some groups rather than others, then "the basis for drawing scientific inference is severely lacking: findings of differences between individuals and groups cannot be unambiguously interpreted" (Horn & McArdle, 1992; p.117).

Establishing model fit for each sub-sample is required before conducting MI analyses (Byrne, Shavelson, & Muthén, 1989). Therefore, we conducted separate Confirmatory Factor Analyses (CFA) for the pooled sample and for each ethnic group, to confirm the proposed

factorial structure of the SPPCC scale. The six-factor model was also tested to compare and see if it confirmed the findings of Rowe et al. (2010).

Since the data were ordered-categorical, a Weighted Least Squares Mean and Variance Adjusted (WLSMV) estimator with theta parameterization was used, both for CFA and MI analyses. The problems associated with treating the ordinal categorical responses as continuous is well-established in the literature (Lubke & Muthén, 2004).

Data fit for the model were assessed by employing multiple criteria (Hair, Anderson, Tatham, & Black, 1998). The chi-square (χ^2) likelihood ratio test, the root mean square error of approximation (RMSEA), Tucker-Lewis index (TLI), comparative fit index (CFI), and the Weighted Root Mean Square Residual (WRMR) were used to test model fit. Acceptable model fit is assessed by a non-significant χ^2 with p degrees of freedom, RMSEA having values less than .08, and CFI and TLI with values greater than .90. Yu (2002) suggested using WRMR instead of Standardized Root Mean Square Residual (SRMR) for categorical outcomes. However, it is important to note that there are no commonly agreed cut-off standards available in the literature for this index and, therefore, it is not recommended as a sole indicator. Additionally, values less than .05 for RMSEA and values at, or above, .95 for CFI and TLI are considered a good fit.

After establishing model to data fit, we performed multi-group confirmatory factor analyses (MG-CFA) which are commonly used to assess invariance of measurement instruments. Following are the recommended steps (Vandenberg & Lance, 2000) for hierarchically nested MI tests:

1. Configural invariance: invariant factor structure across groups.
2. Metric invariance: invariant factor loadings across groups.
3. Scalar invariance: invariant item intercepts (or thresholds) across groups.

At each level, previous invariance tests must be in place. Configural invariance indicates whether or not the scale constructs are conceptualized in the same way by students from different groups. Metric invariance requires the equivalency of measurement units of the scale (latent variable) across groups. In statistical terms, this means that regression slopes (factor loadings) that relate observed variables to latent variables should be identical for the groups. Metric invariance tests the strength of the relationship between items and constructs across groups to see if the responses to items are similar among participants. Scores on the latent variable can still be biased across groups even if the equivalency of measurement units has been established. Therefore, it is necessary to establish scalar invariance, which means that "observed scores are related to the latent scores; that is, individuals who have the same score on the latent construct would obtain the same score on the observed variable, regardless of their group membership" (Milfont & Fischer, 2010, p. 115). Scalar invariance requires that scales of the latent variables have the same measurement unit and the same origin (equivalency of regression intercepts/thresholds) for all groups. (Wu, Li, & Zumbo, 2007). According to Schmitt and Kuljanin (2008), scalar invariance is needed and sufficient to make meaningful comparisons across group means.

Setting the metric of a latent variable is generally accomplished either by fixing the factor variance at one or fixing one of the factor loadings (or referent/marker item) to one, across groups. The latter is recommended in the literature (Stark, Chernyshenko, & Drasgow, 2006) and was utilized in this study, although the selection of a referent item can influence the invariance results (Vandenberg & Lance, 2000), as there is an assumption that the referent indicator is invariant across groups. Therefore, the selection of this indicator should not be at random (Sass, 2011).

To select the most invariant referent items in the SPPCC scale factors, we employed a constrained-baseline (top-down) approach which starts with a most restrictive model (scalar

in this case) which is then compared with models in which single item parameters (loadings and thresholds) are freely estimated across groups. To be able to test all items, the metric of the factors was set by fixing the factor variances at one in the first group and free in other groups.

After identifying the referent items for each SPCC factor, we proceeded with testing the configural, metric, and scalar invariance steps. When evaluating models, factor loadings of the referent items were fixed to 1, allowing factor variances to be freely estimated across groups. A detailed description of the parameter specifications across groups that we employed (Millsap, 2011; Muthén & Muthén, 2013) is provided in Table 1.

--- Insert Table 1 here ---

When non-invariance was found on a scale level, we proceeded with item-level analyses in search of the items that were not functioning equally well across groups. For each item, invariance constraints were relaxed one at a time and the corresponding models were compared, following a Bonferroni adjustment because of the number of comparisons made.

Conventionally, a chi-square difference test ($\Delta\chi^2$) is used to test the difference in fit between unconstrained and constrained models. Due to over-sensitivity of $\Delta\chi^2$ to sample size, Cheung and Rensvold (2002) suggested ΔCFI with absolute values equal to or less than .01 as indicative of adequate MI, which has since become the most frequently used criterion in the literature. Recently, Chen (2007) suggested using a change of .01 and .015 in CFI and RMSEA respectively when the sample size was greater than 300. However, their studies were based on continuous data using the maximum likelihood (ML) estimation method. It is still unknown whether these rules are also applicable to models with categorical indicators estimated with WLSMV. Moreover, traditional $\Delta\chi^2$ cannot be employed for nested models using the WLSMV estimator because WLSMV chi-square values are not distributed as chi-square. Namely, when the WLSMV estimator is used, by employing a two-step procedure

Mplus makes an adjustment to the chi-square (χ^2) and degrees of freedom (df) to obtain a corrected p-value. That is why neither χ^2 nor the df are interpreted in the regular way (Asparouhov & Muthén, 2006). Therefore, model comparisons in this study were made using the DIFFTEST option in Mplus both for scale-level and item-level analyses.

Results

Descriptive Statistics

Missing data were treated using the Expectation Maximization (EM) algorithm, although the percentage of missing cases was trivial (less than 1%). Descriptive statistics are summarized in Table 2. No univariate outliers were found to have an effect on the results. The skewness and kurtosis values indicated that the student responses were distributed fairly normally except for item 4. Rowe at al. (2010) also reported this item not distributing normally. The means and standard deviations of the 26 items ranged from 3.07 to 4.57 and .83 to 1.38, respectively, suggesting that most students indicated either sometimes false/sometimes true or mostly true.

--- Insert Table 2 here ---

Confirmatory Factor Analysis (CFA)

In this study, we utilized separate CFAs with Mplus 7 for the whole (pooled) sample and for each ethnic group.

--- Insert Table 3 here ---

As presented in Table 3, the SPPCC model, consisting of four factors and 26 items, provided acceptable fit for the ethnic groups and whole sample. For this model, all RMSEA values were found to be less than .08 and to have CFI/TLI values greater than .90, which are accepted as the cut-off values for adequate model to data fit (Hu & Bentler, 1999). Our

analyses indicated similar, but slightly higher, factor correlations than those reported by Rowe et al. (2010), and are presented in Table 4.

--- Insert Table 4 here ---

The correlations between SPPCC constructs ranged from .450 to .612 and, therefore, were moderate. Because of known problems of using Cronbach's alpha, especially when the normality assumptions are violated (Teo & Fan, 2013), the linear structural equation modeling (SEM) estimate of reliability, as suggested by Yang and Green (2011), was reported for each factor. The SEM reliability coefficients of scores were consistent with, but slightly lower than, the reliability estimates (Cronbach's alpha) reported by Rowe et al. (2010). Following the analyses, we confirmed that the results supported the four-factor model. CFA analyses and the factor correlations clearly implied that the hypothesized scale structure by Rowe et al. (2010) held for our New Zealand sample.

After confirming the factor structure of the SPPCC scale, we investigated the factor loadings of the scale across sub-samples. The unstandardized and standardized factor loadings are presented in Table 5.

--- Insert Table 5 here ---

All factor loadings were significant. Factor loadings for the pooled sample ranged from .266 to .901. Negatively worded items yielded noticeably lower factor loadings when compared to positively worded items. Having this pattern was deemed acceptable because Chen, Rendina-Gobioff, and Dedrick (2007) demonstrated that young students have difficulty responding appropriately to negatively worded items. Satisfaction (SA) factor items yielded relatively higher loadings than the items from other factors. After establishing a satisfactory-fitting model for each sub-group, we performed MI analyses across ethnicity.

Measurement Invariance (MI) Analysis

We preferred MG-CFA rather than Item Response Theory-based methods because of its better performance with polytomous data (Stark, Chernyshenko, & Drasgow, 2006). Configural, metric, and scalar invariance were tested using the WLSMV estimator, with theta parameterization, due to the categorical nature of the data. Constrained-baseline analyses revealed that items 5, 12, 20, and 23 were the most invariant items for the SPPCC factors. Therefore, those items were chosen as referent items. MI results are presented in Table 6.

--- Insert Table 6 here ---

The results showed that the configural model fitted the data well, indicating that students from different ethnic groups used the same conceptual framework to answer the SPPCC scale items. Results for the configural invariance model supported further investigation of MI. Metric invariance was also established across groups. Scalar invariance results, however, indicated a lack of invariance, DIFFTEST (222) = 347.26, $p < .01$. Both for the configural and metric invariance tests, the number of absolute correlation residuals with values greater than .20 for European, Māori, Pasifika, and Asian samples were 3, 10, 10, and 8 respectively, whereas the values were 3, 9, 11, and 8 for scalar invariance. Almost all of these residuals were found between negatively worded items. The largest correlation residuals were observed between item3-item13 and item8-item13 in each sample for the configural, metric, and scalar tests.

We then attempted to pinpoint the source of this non-invariance at the item level. With the exception of referent items, each item was tested one at a time, for equivalence in separate runs. Since scalar invariance was not achieved, we freed each item's thresholds (except the first threshold) across groups. A chi-square difference test (DIFFTEST) was then applied to see if the difference was statistically significant. If it was significant we could conclude that the item was lacking invariance across groups. A Bonferroni correction

(.05/22) was applied while evaluating the significance level. A summary of the item-level analyses is provided in Table 7.

--- Insert Table 7 here ---

It is evident from Table 7 that items 6, 7, 10, 15, and 19 were not invariant across ethnic groups. Thus, scale-level and item-level MI analyses provided evidence that latent variables accounting for the classroom climate demonstrated partial scalar invariance. Further investigation of the residuals of the intercepts/thresholds for the non-invariant items, revealed that the residuals ranged from -.20 to .31. The largest residuals (.28 and .31) were found to be associated with the fourth thresholds of item 7 and 15 for the Pasifika sample.

We also computed the latent means for the full scalar and partial scalar model, to compare and see if the statistical significance between these models was also practically significant. Non-invariant items (6, 7, 10, 15, and 19) were allowed to have different thresholds for the partial scalar model. The estimated latent means for each ethnic group, for the scalar and partial scalar models, are presented in Table 8.

--- Insert Table 8 here ---

According to Table 8, we concluded that, on average, the latent mean difference between the full scalar and partial scalar models was reasonably similar. We observed, however, a slight difference for the academic competence factor. This was probably due to the smaller number of items in this factor.

Observing relatively lower factor loadings and large correlation residuals between negatively worded items, led us to further investigate the effect of a possible negative wording factor (NWF). Thus, an additional NWF-with items 3, 8, 13, 17, and 24- was modelled to the four-factor SPPCC as a group factor, to account for the variance due to item phrasing.

Then two models were compared: a four-factor SPPCC model (Model 1), and a four-factor SPPCC model with NWF (Model 2), which are presented in Figure 1.

--- Insert Figure 1 here ---

Compared with Model 1, Model 2 provided a better fit for the pooled sample, χ^2 (1855.62, $p < .01$, RMSEA = .05, CFI = .96, TLI = .96, and WRMR = 1.782. Scale level invariance results for Model 2, on the other hand, showed a lack of scalar invariance (DIFFTEST (219) = 350.49, $p < .01$) akin to the four-factor SPPCC (Model 1). At the item level, the same items, except item 19, were not invariant across ethnic groups. Modelling a negative wording factor improved model fit but had negligible effect on the invariance results.

Discussion

The present study adds to the cross-cultural classroom climate research literature, specifically by providing empirical evidence supporting the measurement invariance of the four-factor SPPCC scale across four ethnic groups in New Zealand primary schools. Results of the invariance tests of the SPPCC across four ethnic groups (European, Māori, Pasifika, Asian) indicated that the SPPCC represented the same four classroom climate factors for each group (configural invariance). Full scalar invariance and comparability of the SPPCC constructs across the different groups was not supported. Only partial scalar invariance was achieved because of a lack of invariance in the thresholds of five items (items 6, 7, 10, 15 and 19). In other words, these items displayed differential item functioning and thus likely meant something different to students with European and Pasifika ethnic backgrounds. Nevertheless, latent mean comparison between the full and partial scalar models indicated that the scalar noninvariance may not be practically significant.

This study follows Rowe et al's (2010) recommendations that the SPPCC should be subject to further empirical testing, and with younger elementary school and middle school students in order to "better understand the potential role of development on students' own

perceptions of their classroom climate" (p. 874). The results have provided evidence that the four-factor model proposed by Rowe et al. (2010) could be operationalized in the SPPCC, and equally applied across the four groups of New Zealand primary school students. This study has several important implications. First, results of the invariance tests verify the generalizability of the SPPCC to different cultural groups within a multicultural society, such as New Zealand. The initial study (Rowe et al., 2010) had not considered the cultural applicability of the SPCC. Second, there is support for using this scale to assess the class climate with elementary school students. Third, by establishing the measurement invariance of the SPPCC in the New Zealand context, future studies could further investigate the use of the SPCC in other cultural contexts, for example, within an Asian setting or among other indigenous groups. Many studies employ measures developed in a western context with other cultural groups, without regard for how different groups might respond (King & McInerney, 2014). It is important to establish cultural invariance before measures are employed in contexts in which they were not developed because, otherwise, the measurement of student beliefs may not be reflective of their actual beliefs (Zusho & Clayton, 2011). Finally, this study addressed the lack of advanced use of statistical techniques in classroom climate research by using tests of measurement invariance , to offer evidence for the structure of the measure with elementary school students.

Limitations of the Study

Several limitations should be mentioned. First, despite steps taken to ensure that the classrooms where data were collected were non-threatening, by ensuring teacher absence, a possible bias may exist in students' responses. Students' responses may have been influenced by their peers' presence and they may have responded more positively than their actual perceptions, a situation known as social desirability. One way of dealing with social desirability would be to allow respondents to give their responses confidentially, for example,

by using an online questionnaire. Second, as factor analysis is sample dependent, caution should be exercised when applying the findings to other samples with different profiles such as age, culture, and socioeconomic status. Third, the use of negative items in the SPPCC among young children may have increased the cognitive complexity in the items, resulting in some disruption in the factor structure. Therefore, despite the support provided by this study of the suitability of the SPPCC for use among elementary school students, further studies aimed at increasing empirical support could be conducted in other multicultural societies, for example, in Europe. This would facilitate a broader understanding of the variables associated with a positive class climate, as perceived by students from a range of different groups. Studies in other contexts could eventually lead to researchers being able to establish a nomological network of variables associated with the class climate.

Contributions of the Study

The SPPCC is a relatively simple instrument to administer and to score, and given the strong association that has been found between positive teacher-student relationships and achievement (Hattie, 2009; $d = 0.72$), it would seem worthwhile for teachers to come to understand more about the psychosocial aspects of their classrooms, rather than exclusively concentrating on pedagogical methods. The SPPCC could potentially be employed by individual teachers or schools, in order to assess the class climate within specific classrooms or across groups. At the teacher-level, this would enable teachers to determine areas of the class climate where they were successful, as well as those that needed improvement. Teachers could also evaluate their own success in relation to various groups, since the SPPCC enables evaluation of both teacher-student and student-student relationships. At the whole school level, information from the SPPCC would enable principals to ascertain any teachers needing help in building a positive class climate. Again, principals could also assess school-wide success in building equitable and positive relationships with different ethnic

groups. This could be important, particularly in New Zealand, where schools are required to report on how well they are serving Māori students. Having teachers and schools become more aware of the importance of the class climate for student achievement is clearly worthwhile. Although there have been many pedagogical interventions over the years, in many countries, aimed at improving the achievement of ethnic minority groups, there have been few gains of note (Author, 2015). It may well be that a focus on enhancing student perceptions at the psychosocial level, through interventions aimed at improving teacher, student, class, and school relationships, may offer a more positive way forward, in the quest to increase student achievement and the life chances of the current underachieving groups.

References

Anderman, L. H., Patrick, H., Hruda, L. Z., & Linnenbrink, E. A. (2002). Observing classroom goal structures to clarify and expand goal theory. In C. Midgley (Ed.), *Goals, goal structures, and patterns of adaptive learning* (pp. 243–294). Mahwah, NJ: Lawrence Erlbaum.

Asparouhov, T. & Muthén, B. O. (2006) Robust Chi Square Difference Testing with Mean and Variance Adjusted Test Statistics. Mplus Web Notes: No. 10. http://statmodel.com/download/webnotes/webnote10.pdf

Author (2015). Blinded for review

Author & colleague (2011). Blinded for review

Babad, E. (2009). *The social psychology of the classroom*. New York: Routledge.

Bandura, A. (1993). Perceived self-efficacy in cognitive development and functioning. *Educational Psychologist, 28*, 117–148.

Bishop, R., & Berryman, M. (2006). *Culture speaks: Cultural relationships and classroom learning*. Wellington, New Zealand: Huia.

Brophy-Herb, H. E., Lee, R. E., Nievar, M. A., & Stollak, G. (2007). Preschoolers' social competence: Relations to family characteristics, teacher behaviours, and classroom climate *Journal of Applied Developmental Psychology, 28*, 134–148.

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

Chen, Y-H., Rendina-Gobioff, G., & Dedrick, R. F. (2007, November). *Detecting effects of positively and negatively worded items on a self-concept scale for third and sixth grade elementary students.* Paper presented at the annual meeting of the Florida Educational Research Association, Tampa, FL.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233–255.

Deci, E., & Ryan, R. (Eds.). (2002). *Handbook of self-determination research*. Rochester, NY: University of Rochester Press.

DiStefano, C., & Molt, R. W. (2006). Further investigationg method effects associated with negatively worded items on self-report surveys, *Structural Equation Modeling:A Multidisciplinary Journal, 13*, 440–464.

Ennis, C. D. (1998). Shared expectations: Creating a joint vision for urban schools. In J. Brophy (Ed.) *Advances in research on teaching: Expectations in the classroom,* 151–182. Greenwich, CT: JAI.

Fraser, B. J. (1986). *Classroom environment*. Abingdon, Oxon, UK: Routledge.

Fraser, B. J. (1998). Classroom environment instruments: Development, validity, and applications. *Learning Environments Research, 1*, 7–33.

Fraser, B. J., Anderson, G. J., & Walberg, H. J. (1982). *Assessment of learning environments: Manual for Learning Environment Inventory (LEI) and My Class Inventory (MCI)*. Perth: Western Australian Institute of Technology.

Fraser, B. J., Giddings, G. J. , & McRobbie, C. J. (1995). Evolution and validation of a personal form of an instrument for assessing science laboratory classroom environments. *Journal of Research in Science Teaching, 32*, 399–422.

Freeman, K. E., Gutman, L. M., & Midgley, C. (2002). Can achievement goal theory enhance our understanding of the motivation and performance of African American young adolescents? . In C. Midgley (Ed.), *Goals, goal structures and patterns of adaptive learning* (pp. 176–204). Mahwah, NJ: Lawrence Erlbaum.

Furrer, C., & Skinner, E. (2003). Sense of relatedness as a factor in children's academic engagement and performance. *Journal of Educational Psychology, 95*, 148–162.

Hair, J. F., Anderson, R. E., Tatham, R. L. & Black, W. (1998). *Multivariate data analysis* (5th ed.). New York: Prentice Hall.

Hamre, B. K., & Pianta, R. C. (2001). Early teacher-child relationships and the trajectory of children's school outcomes through eighth grade. *Child Development, 72*, 625–638.

Harter, S. (1985). *Manual for the self-perception profile for children*. Denver, CO: University of Denver.

Hattie, J. A. C. (2008). Narrow the gap, fix the tail, or close the curves: The power of words. In Rubie-Davies & Rawlinson (Eds.) *Challenging thinking about teaching and learning,* 19-23. New York, NY: Nova.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses.* London, UK: Routledge.Hong, S., Malik, M. L., & Lee, M. K. (2003). Testing configural, metric, scalar and latent mean invariance across gender in sociotropy and autonomy using nonwestern sample. *Educational and Psychology Measurement*, *63*, 636–653.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*, 117–144.

Hornstra, L., van der Veen, I., Peetsma, T., & Volman, M. (2013). Developments in motivation and achievement during primary school: A longitudinal study on group-specific differences. *Learning and Individual Differences, 23*, 195–204.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.

Huebner, E. S. (1994). Preliminary development and validation of a multidimensional life satisfaction scale for children. *Psychological Assessment, 6*, 149–158.

Johnson, D. W., Johnson, R. T., & Anderson, D. (1983). Social interdependence and classroom climate. *Journal of Psychology, 114*, 135–142.

King, R. B., & McInerney, D. M. (2014). Culture's consequences on student motivation: Capturing universality and variability through personal investment theory. *Educational Psychologist, 49,* 175–198. doi: 10.1080/00461520.2014.926813

Lubke, G. H., & Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling, 11*(4), 514–534.

McKown, C., & Weinstein, R. S. (2008). Teacher expectations, classroom context and the achievement gap. *Journal of School Psychology, 46*, 235–261.

Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research, 3*, 111–121.

Millsap, R. E. (2012). *Statistical approaches to measurement invariance*. New York: Taylor and Francis Group.

Moos, R. H. (1979). *Evaluating educational environments: Procedures, measures, findings and policy implications*. San Francisco: Jossey Bass.

Moos, R. H., & Trickett, E. (1987). *Classroom environment scale manual* (2nd ed.). Palo Alto, CA: Consulting Psychologists Press.

Muthén, L. K. & Muthén, B. O. (1998-2010). *Mplus user's guide*. Sixth edition. Los Angeles, CA: Muthén & Muthén

Muthén L. K., Muthén B. O. (2013). *Version 7.1 Mplus language addendum*. Available online at: http://www.statmodel.com/ugexcerpts.shtml.

Nunn, L. M. (2011). Classrooms as racialized spaces: Dynamics of collaboration, tension, and student attitudes in urban and suburban high schools. *Urban Education, 46,* 1226–1255. doi: 10.1177/0042085911413146

Peterson, P. L., & Walberg, H. J. (1979). *Research on teaching: Concepts, findings, and implications*. Berkeley, CA: McCutchan.

Raykov, T., & Marcoulides, G. A. (2008). *An Introduction to Applied Multivariate Analysis.* New York: Taylor & Francis.

Roeser, R. W., Eccles, J. S., & Sameroff, A. J. . (2000). School as a context of early adolescents' academic and social-emotional development: A summary of research findings. *Elementary School Journal, 100*, 443–471.

Rowe, E. W., Kim, S. , Baker, J. A., Kamphaus, R. W. , & Horne, A. M. (2010). Student personal perception of classroom climate: Exploratory and confirmatory factor analyses. *Educational and Psychological Measurement, 70*, 858–879.

Sass, D. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *Journal of Psychoeducational Assessment*, *29*(4), 347-363.

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review, 18*, 210–222.

Schweinle, A., & Mims, G. A. (2009). Mathematics self-efficacy: Stereotype threat versus resilience. *Social Psychology of Education: An International Journal, 12*, 501–514.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*, 1292–1306.

Statistics New Zealand (2010). http://www.stats.govt.nz/browse_for_stats/population/estimates_and_projections/demographic-trends-2010.aspx

Statistics New Zealand (2013). http://www.stats.govt.nz/Census/2013-census.aspx

Taylor, P. B., Fraser, B. J., & Fisher, D. (1997). Monitoring constructivist classroom learning environments. *International Journal of Educational Research, 27*, 293–302.

Teo, T., & Fan, X. (2013). Coefficient alpha and beyond: Issues and alternatives for educational research. *The Asia-Pacific Education Research, 22*, 209–213.

Urdan, T., & Schoenfelder, E. (2006). Classroom effects on student motivation: Goal structures, social relationships, and competence beliefs. *Journal of School Psychology, 44*, 331–349.

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,* 4–69.

Walberg, H. J. (1979). *Educational environments and effects: Evaluation policy and productivity*. Berkeley, CA: McCutchan.

Weinstein, R. S. (2002). *Reaching higher: The power of expectations in schooling*. Cambridge, MA: Harvard University Press.

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research & Evaluation, 12*(3), Retrieved from: http://pareonline.net/getvn.asp?v=12&n=13.

Yang, Y., & Green, S. B. (2011). Coefficient alpha: A reliability coefficient for the 21st century?. *Journal of Psychoeducational Assessment, 29*(4). 377-392.

Yu, C. Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes* (Doctoral dissertation, University of California Los Angeles).

Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language testing, 20*(2), 136-147.

Zusho, A., & Clayton, K. (2011). Culturalizing achievement goal theory and research. *Educational Psychologist*, *46*, 239–260. doi: 10.1080/00461520.2011.614526

Table 1.

Scaling latent variables

	Configural	Metric	Scalar
Loadings	Free	Equal	Equal
Thresholds	Free	First threshold of each item is equal Second threshold of referent item is equal	Equal
Residual variances	Fixed at 1	Fixed at 1 in first group Free in other groups	Fixed at 1 in first group Free in other groups
Factor means	Fixed at 0	Fixed at 0 in first group Free in other groups	Fixed at 0 in first group Free in other groups
Factor variances	Free	Free	Free

Table 2.

Descriptive Statistics for all SPPCC scale items

Factors/Items	*M*	*SD*	Skewness	Kurtosis
TEACHER SUPPORT (TS)				
1. My teacher cares about how much I learn.	4.32	0.96	-1.44	1.71
2. My teacher likes to see my work.	4.14	0.95	-0.98	0.59
3.* My teacher does not like to help me learn.	4.30	1.18	-1.67	1.72
4. My teacher wants me to do my best school work.	4.57	0.83	-2.25	5.21
5. My teacher really cares about me.	4.15	1.00	-1.04	0.58
6. My teacher thinks it is important for her/him to be my friend.	3.70	1.16	-0.57	-0.43
7. My teacher likes me as much as he/she likes other students in the class.	3.86	1.28	-0.88	-0.30
8.* My teacher does not care my feelings.	4.22	1.18	-1.44	1.00
PEER SUPPORT (PS)				
9. The kids in my class want me to do my best schoolwork.	3.46	1.22	-0.39	-0.64
10. The kids in my class like to help me learn.	3.37	1.24	-0.27	-0.78
11. The kids in this class care about how much I learn.	3.07	1.25	-0.01	-0.87
12. The kids in this class want me to come to class every day.	3.49	1.22	-0.39	-0.71
13.* In this class, other students do not think it is important to be my friend	3.56	1.28	-0.49	-0.78
14. In this class, other students like me the way I am.	3.98	1.12	-0.95	0.20
15. In this class, other students care about my feelings.	3.70	1.14	-0.61	-0.32
16. In this class, other students really care about me.	3.53	1.17	-0.40	-0.57
ACADEMIC COMPETENCE (AC)				
17.* I am not very good at my schoolwork.	3.66	1.29	-0.59	-0.73
18. I am smart enough to do my schoolwork.	4.02	1.10	-0.99	0.26
19. I do very well at my schoolwork.	3.91	1.01	-0.71	0.05
20. I can work out the answers to schoolwork.	3.93	1.02	-0.74	0.12
SATISFACTION (SA)				
21. I look forward to going to school.	4.01	1.19	-1.02	0.08
22. I like being in school.	4.05	1.16	-1.07	0.24
23. School is interesting.	4.04	1.15	-1.06	0.27
24.* I wish I didn't have to go to school.	3.84	1.38	-0.89	-0.52
25. There are many things about school that I like.	4.20	1.06	-1.24	0.82
26. I enjoy school activities.	4.30	0.99	-1.35	1.19

*: Negatively worded items

Table 3.

CFA results for pooled sample and ethnic groups

Models	χ^2	*df*	p	RMSEA (90 % CI)	CFI	TLI	WRMR
Four-factor model							
Pooled	3043.58	293	.00	.07 (.07, .07)	.93	.93	2.45
European	1443.31	293	.00	.07 (.06, .07)	.95	.94	1.68
Māori	856.39	293	.00	.07 (.07, .08)	.93	.92	1.40
Pasifika	715.82	293	.00	.07 (.06, .07)	.92	.91	1.26
Asian	669.15	293	.00	.07 (.06, .07)	.94	.94	1.23
Six-factor model							
Pooled	2570.38	284	.00	.07 (.06, .07)	.94	.94	2.18
European	1184.94	284	.00	.06 (.06, .06)	.96	.95	1.47
Māori	777.48	284	.00	.07 (.06, .08)	.94	.93	1.29
Pasifika	672.54	284	.00	.07 (.06, .07)	.93	.91	1.19
Asian	611.57	284	.00	.06 (.06, .07)	.95	.94	1.14

Note. RMSEA, Root Mean Square Error of Approximation; CFI, Comparative Fit Index; TLI, Tucker-Lewis Index; WRMR, Weighted Root Mean Square Residual.

Table 4.

Factor correlations and reliabilities for pooled sample

	TS	PS	AC	SA
Teacher Support (TS)	-			
Peer Support (PS)	.61	-		
Academic Competence (AC)	.60	.45	-	
Satisfaction (SA)	.57	.54	.58	-
SEM Estimate of Reliability	.70	.77	.74	.87

Note. TS; Teacher Support, PS; Peer Support, AC; Academic Competence, SA; Satisfaction

Table 5.

Standardized and unstandardized factor loadings&standard errors (in parentheses) of SPPCC

Items	Pooled	European	Māori	Pasifika	Asian
1	.76 (1.00; .00)	.75 (1.00; .00)	.75 (1.00; .00)	.79 (1.00; .00)	.75 (1.00; .00)
2	.69 (.81; .05)	.69 (.85; .07)	.68 (.81; .13)	.73 (.82; .13)	.65 (.77; .13)
3	.45 (.44; .04)	.46 (.46; .06)	.33 (.31; .07)	.53 (.49; .09)	.51 (.54; .11)
4	.62 (.68; .05)	.64 (.74; .07)	.57 (.61; .10)	.65 (.67; .13)	.60 (.67; .12)
5	.80 (1.15; .07)	.81 (1.22; .11)	.84 (1.35; .21)	.76 (.90; .13)	.76 (1.05; .16)
6	.63 (.70; .05)	.63 (.73; .07)	.62 (.69; .11)	.63 (.62; .11)	.59 (.66; .10)
7	.47 (.46; .03)	.55 (.58; .06)	.36 (.34; .06)	.41 (.35; .07)	.52 (.55; .11)
8	.50 (.50; .04)	.55 (.58; .06)	.50 (.51; .08)	.44 (.38; .07)	.48 (.49; .08)
9	.65 (1.00; .00)	.69 (1.00; .00)	.55 (1.00; .00)	.71 (1.00; .00)	.60 (1.00; .00)
10	.72 (1.20; .06)	.72 (1.09; .08)	.66 (1.30; .19)	.82 (1.47; .20)	.65 (1.15; .17)
11	.73 (1.24; .06)	.72 (1.08; .08)	.72 (1.60; .22)	.79 (1.28; .17)	.70 (1.31; .18)
12	.70 (1.16; .06)	.72 (1.09; .08)	.68 (1.40; .20)	.68 (.94; .12)	.72 (1.38; .20)
13	.27 (.32; .03)	.31 (.34; .05)	.19 (.30; .09)	.17 (.19; .06)	.36 (.52; .10)
14	.71 (1.19; .07)	.74 (1.15; .09)	.71 (1.53; .23)	.63 (.83; .13)	.76 (1.60; .25)
15	.79 (1.51; .09)	.82 (1.50; .12)	.75 (1.72; .23)	.76 (1.16; .15)	.83 (2.02; .27)
16	.79 (1.50; .08)	.85 (1.69; .12)	.68 (1.43; .22)	.75 (1.14; .17)	.77 (1.62; .22)
17	.51 (1.00; .00)	.56 (1.00; .00)	.39 (1.00; .00)	.44 (1.00; .00)	.64 (1.00; .00)
18	.78 (2.06; .15)	.79 (1.92; .20)	.78 (2.95; .55)	.75 (2.33; .45)	.78 (1.51; .26)
19	.84 (2.59; .21)	.84 (2.31; ,25)	.86 (3.69; .82)	.79 (2.62; .50)	.84 (1.85; .35)
20	.76 (1.98; .15)	.78 (1.83; .19)	.74 (2.60; .46)	.73 (2.18; .41)	.78 (1.47; .22)
21	.87 (1.00; .00)	.88 (1.00; .00)	.91 (1.00; .00)	.81 (1.00; .00)	.81 (1.00; .00)
22	.90 (1.18; .07)	.91 (1.17; .09)	.90 (.88; .13)	.88 (1.35; .18)	.92 (1.71; .30)
23	.85 (.90; .05)	.84 (.86; .06)	.85 (.72; .10)	.83 (1.07; .14)	.86 (1.21; .17)
24	.68 (.52; .03)	.69 (.51; .04)	.67 (.40; .05)	.64 (.61; .08)	.72 (.77; .12)
25	.82 (.80; .04)	.82 (.79; .06)	.78 (.55; .07)	.83 (1.08; .15)	.81 (1.02; .16)
26	.81 (.79; .04)	.81 (.75; .05)	.85 (.71; .10)	.82 (1.03; .14)	.74 (.82; .13)

Table 6.

Measurement invariance results at the scale level

	χ^2	*df*	RMSEA (90% CI)	CFI	WRMR	$\Delta\chi^2$ (DIFFTEST)	Δdf	p
Configural	3530.89	1172	.07 (.06, .07)	.95	2.81	-	-	-
Metric	3369.98	1238	.06 (.06, .06)	.95	2.48	56.21	66	.80
Scalar	3506.90	1460	.06 (.05, .06)	.95	2,97	347.26	222	.00

Note. RMSEA, Root Mean Square Error of Approximation; CFI, Comparative Fit Index; WRMR, Weighted Root Mean Square Residual.

Table 7.

Measurement invariance results at the item level

Factor	Items	$\Delta\chi^2$ (DIFFTEST)	p
	1	7.10	.63
	2	19.86	.02
	3	5.09	.83
	4	6.07	.73
TS	5*	-	-
	6	**28.31**	< .01
	7	**42.40**	< .01
	8	12.06	.21
	9	10.54	.31
	10	**28.08**	< .01
	11	16.35	.06
PS	12*	-	-
	13	8.76	.46
	14	11.59	.24
	15	**51.64**	< .01
	16	19.24	.02
	17	20.34	.02
AC	18	16.11	.06
	19	**26.30**	< .01
	20*	-	-
	21	13.97	.12
	22	8.72	.73
SA	23*	-	-
	24	20.42	.02
	25	11.25	.26
	26	17.87	.04

Note: Degrees of freedom is 9 for all models. Significant DIFFTEST values were highlighted in bold after Bonferroni correction ($\alpha = .05/22 = .002$).
*: Referent items

Table 8.

Estimated latent means for ethnic groups

Models	Scalar Model				Partial Scalar Model			
	TS	PS	AC	SA	TS	PS	AC	SA
European	.00	.00	.00	.00	.00	.00	.00	.00
Māori	-.04	.13	.00	.51	-.03	.14	-.05	.50
Pasifika	.09	.28	-.11	.99	.11	.28	-.21	.99
Asian	-.16	-.06	.16	.50	-.19	-.08	.18	.50

Note. TS; Teacher Support, PS; Peer Support, AC; Academic Competence, SA; Satisfaction

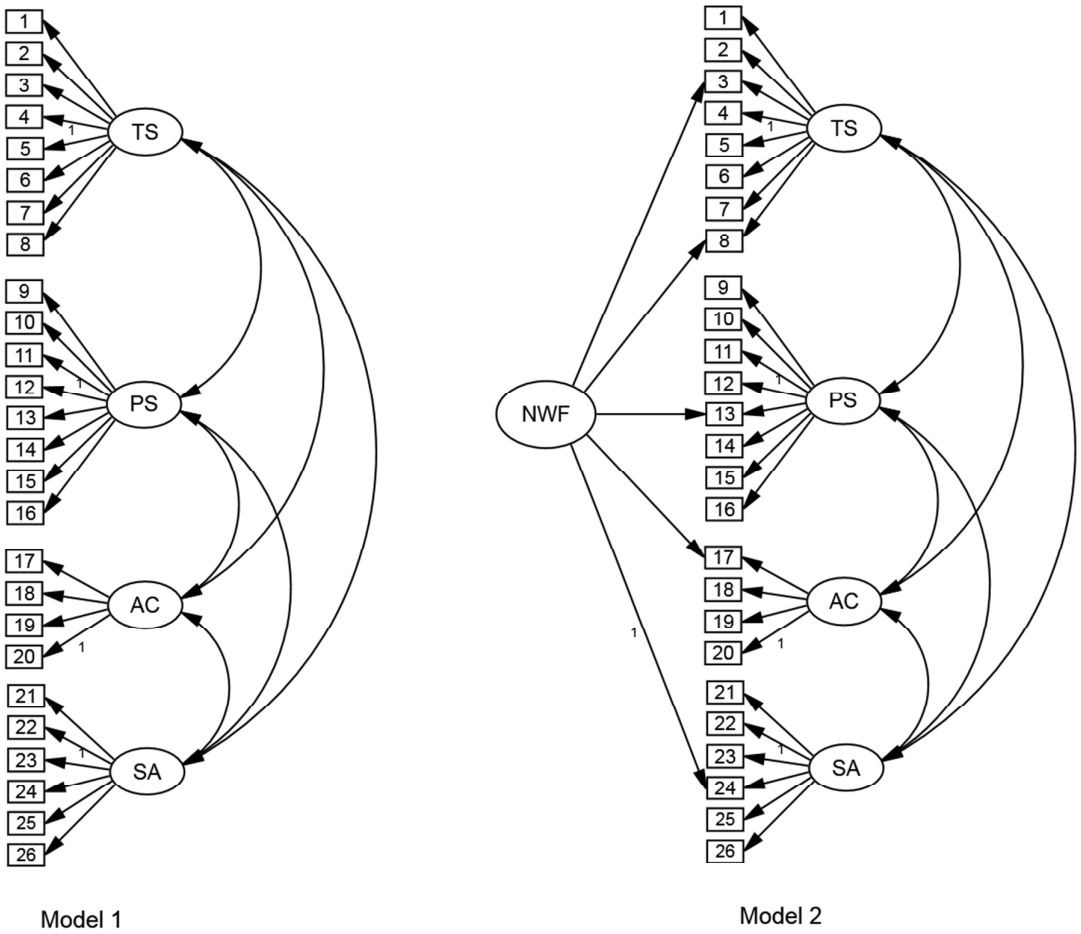

Figure 1. Alternative models of SPPCC.

Title: Assessing measurement invariance of the Student Personal Perception of Classroom Climate (SPPCC) across different ethnic groups (JPA-14-0024.R5)

Editor's Comment	Author's Response
I reviewed your manuscript and believe some minor revisions are needed prior to publication. First, in the abstract, I'm not sure what is meant by "multiple-group CFA contrasting the SPPCC with four ethnic samples". The invariance tests you conducted involved constraining parameters such as factor loadings and intercepts to be equal across groups. Thus, the SPPCC isn't being contrasted with ethnic groups, rather you are testing the equivalence of model parameters across ethnic groups.	This part of the abstract has been amended.
Second, there are a couple of paragraphs that begin with the word "nevertheless". The word nevertheless suggests continuation of the previous thought, while paragraph breaks reflect changes in topics. Please revise accordingly.	These have been amended.
Third, on pages 9 and 11 it is stated that the current study aimed to examine classroom climate perceptions across four different ethnic groups. This seems inaccurate based on my read of the manuscript. You examined invariance, but you do not specifically report and discuss comparisons of negative perceptions across groups. Therefore, this text should be deleted.	This portion of the text has been amended.
Fourth, on page 12, first paragraph: The word "the" is missing before Student Personal Perception of Classroom Climate.	The word 'the' has been added.
Fifth, on page 15, the sentence "Further, establishing scalar invariance indicates that observed scores are related to the latent scores" doesn't seem to follow from the previous sentences. Please revise. For example: Metric invariance tests the strength of the relationship between items and constructs across groups to see if the responses to items are similar among participants. Scores on the latent variable can still be biased across groups even if the equivalency of measurement units as been established. Therefore, it is necessary to establish scalar invariance, which means that observed scores are related to the latent scores; that is, individuals who have the same score on the latent construct would obtain the same score on the observed variable, regardless of their group membership" (Milfont & Fischer, 2010, p. 115).	The wording suggested by the author has been used.
Sixth, on page 18 SEM reliability estimates are discussed, but no reference is made to Table 4 where these values are reported.	Table 4 reports the Factor correlations and reliabilities for pooled sample. Reference to Table is made in the sentence in page 18, "Our analyses indicated similar, but slightly higher, factor correlations than those reported by Rowe et al. (2010), and are presented in Table 4."
Seventh, the conclusion that this study confirm a four factor structure is a stretch. Aside from examining a six factor alternative, no tests of alternative models were conductive. The obtained fit statistics do not suggest excellent fit. Therefore, it seems more apt to state something such as "the results support the four-factor model".	The wording has been changed in line with the Editor's suggestion.
Eighth, on page 22 please revise the following: "Among the contributions this study makes are first". For example, you might state "This study has several important implications. First,……." Also, I'm not sure what is meant by "generalizability of the SPPCC to a multicultural society, such	This wording has been changed in line with the Editor's suggestion. This statement has been clarified.

as New Zealand". Please revise this statement to be more specific.	
Finally, please revise the phrase "further empirical support for this instrument could be made in future studies by comparing how well it tests invariance in other cultural context" (p.22). Suggesting that empirical support could be made gives an impression of bias toward finding the results you desire. Also, the SPPCC cannot test invariance.	This wording has been revised.