Configural and metric invariance of the teacher classroom behavioral climate scale across Finland and Greece

Emmi Pelkonen

Master's Thesis in Special Education
Spring Term 2023
Department of Education
University of Jyväskylä

Tiivistelmä

Pelkonen, Emmi. 2023. Configural and metric invariance of the teacher classroom behavioral climate scale across Finland and Greece. Erityispedagogiikan pro gradu –tutkielma. Jyväskylän yliopisto. Kasvatustieteiden laitos. 35 sivua.

Tämän tutkimuksen tarkoituksena oli tarkastella suomalaiseen työrauhatutkimukseen kehitetyn opettajan arviota luokan työrauhasta mittaavan mittarin faktorirakenteen ja -latausten invarianssia Suomen ja Kreikan aineistojen välillä konfirmatorista faktorimallinnusta hyödyntäen.

Mittaria on käytetty useissa Euroopan maissa, mutta mittarin faktorimallin invarianssia eri ryhmien välillä ei ole tarkasteltu aikaisemmin. Mittausinvarianssin tutkiminen osana mittarin rakennevaliditeettia on tärkeää, jotta voidaan varmistua siitä, että tutkittava ilmiö on samanlainen eri konteksteissa.

Suomen aineisto on ProKoulu-tutkimuksen alkumittausaineistosta vuodelta 2013 ja Kreikan aineisto taas Erasmus SWPBS -hankkeen alkumittauksesta vuodelta 2019. Tämän tutkimuksen data koostui 694 opettajan työrauha-arviosta 70 eri peruskoulusta Suomesta ja 269 arviosta 30 eri peruskoulusta Kreikasta. Analyyseissä käytettiin Mplus version 8-tilasto-ohjelmaa.

Ensin työrauhan teoriaan perustuvan nelifaktorimallin sopivuus testattiin maakohtaisesti. Sitten tehtiin moniryhmämallinnus faktorirakenteen invarianssin testaamiseksi. Lopuksi moniryhmämallissa faktorilataukset asetettiin invarianteiksi maiden välillä.

Tutkimuksen tulokset tukevat opettajan luokan työrauhamittarin teoriaan perustuva nelifaktorirakenteen sopimista Suomen ja Kreikan otoksiin. Tulosten perusteella voidaan sanoa mittarin faktorirakenteen ja muuttujien faktorilatausten olevan invariantteja maiden välillä. Tulokset antavat positiivisen kuvan teoriaan perustuvan työrauhamallin toimivuudesta ja sen yleistettävyydestä Suomen ja Kreikan välillä.

Avainsanat: työrauha, mittausinvarianssi, rakennevaliditeetti, konfirmatorinen faktorianalyysi

ABSTRACT

Pelkonen, Emmi. 2023. Configural and metric invariance of the teacher classroom behavioral climate scale across Finland and Greece. Master's Thesis in Special Education. University of Jyväskylä.Department of Education. 35 pages.

Examining validity in different populations is important to ensure that the measurement tool is appropriate and accurate for use with those populations. Classroom behavioral climate (CBC) scale was developed in Finland to measure the behavioral climate specifically in the classroom context. The scale has been used in several countries but measurement equivalence between populations has not been examined before. Purpose of this study is to examine the factor structure and configural and metric invariance of the teachers' CBC scale in Finland and Greece using confirmatory factor analysis.

The Finnish data is from baseline measurement of ProKoulu project from 2013 and the Greek data is from baseline measurement of Erasmus SWPBS project from 2019. Data consists of 694 teacher evaluations of CBC from 70 schools from Finland and 297 evaluations from 30 schools from Greece. Analyses were made by using Mplus Statistical Package Version 8.

First, baseline models for each country were identified. Then, the models were entered into a multiple group analysis to test for configural invariance. Lastly, factor loadings were set to be invariant across the country models to investigate the metric invariance of the scale.

The results showed that the theorized four-factor structure fit both country models and support for both configural and metric invariance was established. These results give a positive outlook on the theorized four factor model of CBC and the use of the scale in Finnish and Greek primary school contexts.

Keywords: classroom behavioral climate, configural invariance, metric invariance, construct validity, confirmatory factor analysis

TABLE OF CONTENTS

1 INTRODUCTION 1

1.1 Behavior and discipline at school 2

1.2 Classroom behavioral climate 4

1.3 Construct validity 5

1.4 Measurement invariance 6

1.5 Research questions 7

2 RESEARCH METHODS 8

2.1 Research context 8

2.2 Data and participants 10

2.3 Data Collection 12

2.4 Data Analysis 14

2.5 Ethical Solutions 16

3 RESULTS 18

3.1 Factor structure of the classroom behavioral climate scale 18

3.2 Configural invariance 20

3.3 Metric invariance 20

4 DISCUSSION 21

REFERENCES 26

1 INTRODUCTION

Validity is fundamental in developing and evaluating tests and measurements in quantitative research. In the social and behavioral sciences different phenomena concerning human behavior are often assessed with self-report questionnaires consisting of items that are developed to assess an underlying construct (Schoot et al., 2012). These questionnaires or scales typically aim to follow individuals over time or compare groups and, for these comparisons to be valid, the scale should measure the construct with the same structure across different contexts (Schoot, et al., 2012; Dimitrov, 2010). Examining measurement validity is important because it ensures that the measurement instrument is accurately measuring the construct it is intended to measure. In other words, it verifies that the results obtained from the tool are meaningful and reliable.

A key issue in educational settings is the degree to which validity evidence based on test-criterion relations can be generalized to new contexts without further study of validity in that new situation (AERA, APA, & NCME, 2010). Examining measurement validity in different contexts or populations is important because the validity of a measurement tool may not generalize across time or different groups of people. Different populations may have different experiences, beliefs, values, and behaviors, which can affect how they respond to a measurement tool. Therefore, examining validity in different populations is important to ensure that the measurement tool is appropriate and accurate for use with those populations.

Classroom behavioral climate (CBC) scale examined in this study has been developed in Finland for the purpose of measuring behavioral climate specifically in the classroom context (see Närhi et al., 2014; Närhi et al., 2017). The scale has been used in studies in Finland (e.g., Närhi et al., 2014; Närhi et al., 2017) and after the piloting of the scale in Finland it has been adapted to other languages and used in other countries. Shortened version of the student CBC scale was used in Germany (Hoffmann et al. 2018) and full student and teacher scales were used in Cyprus, Greece and Romania during the "Building School-Wide Inclusive, Positive and Equitable Learning Environments Through A Systems-Change Approach" (SWPBS) ERASMUS+ Key Action 3 Policy Experimentation program or Erasmus SWPBS project. However, the scale has not yet been tested for its measurement equivalence between different populations.

Examining the scale's factor structure and measurement invariance is crucial in terms of the scale's validation and future utilization. When comparisons among groups on an underlying construct are made, in this case on classroom behavioral climate, it is important to ensure that the assessment instrument is operating in the same way and measures same constructs with the same structure across different groups and contexts (Schoot et al., 2012; Dimitrov, 2010).

The purpose of this study is to examine the configural and metric invariance of the teacher CBC scale between Finnish and Greek samples using confirmatory factor models. The Finnish data is from baseline measurement of ProKoulu project from 2013 and the Greek data is from the baseline measurement of SWPBS Erasmus project from 2019.

1.1 Behavior and discipline at school

Discipline problems are a prevalent issue in schools across the world, affecting both students and teachers. Behavior problems in school have been found to be a risk for student academic achievement (e.g., Finn et al., 2008; Wagner et al., 2005; Hinshaw, 1992), especially when student also has learning difficulties (Algozzine et al., 2011). Problematic behavior also strains students' social relationships - both teacher-student relationship (Nurmi, 2012) and peer relationships (Bollmer, et. al, 2005). Behavior problems among school aged children have been directly and indirectly, through academic and social problems, linked with social and emotional difficulties later in life (e.g., Karakus et al. 2012; Schaeffer et al. 2006).

For teachers, discipline problems in the classroom have been found to be a major cause of increased work-related stress and reduced well-being (Klassen and Chiu, 2010; Hakanen et al., 2006; Boyle et al., 1995). Similarly, self-efficacy in behavior management has been found to have a connection to teachers' job satisfaction and burnout (Malinen & Savolainen, 2016). Overall, positive school climate and learning environment is associated with fewer behavioral problems (Thapa et al., 2013).

Because of the widely perceived importance of behavior in the school context, behavior and phenomena linked to it have been researched extensively in social and behavioral sciences. Consequently, behavior phenomena have been defined and measured in number of ways. Often disruptive behavior of individual students is focused on (e.g., Spilt & Koomen, 2009), but discipline can be seen as part of the wider learning climate of a school or a classroom.

Disciplinary climate is a widely used term that focuses on the ability of the teacher on keeping an orderly classroom (Cheema & Kisantas, 2014; Moos, 1979). In the PISA study (OECD, 2019b) disciplinary climate is measured by the extent to which students miss learning opportunities due to disruptive behavior in the classroom. Disciplinary climate was found to vary widely across countries. It was also noted that variation across schools was large, and, in many countries, the disciplinary problems were highly concentrated in some schools. (OECD, 2019b, 67.) Previous research has also noted that there are variations in disciplinary climate between classrooms (Holopainen et al., 2009) and teachers' strategies on behavior management have an important role establishing a positive disciplinary climate in the classroom (Oliver et al., 2011).

Behavioral climate can also be approached via the identification of desirable and the prevention of undesirable student behaviors (Hochweber et al., 2014). Undesirable behavior, disruptive behavior or misbehavior has also been defined in number of ways. Charles (2005) defines misbehavior in school as behavior that violates class rules, demeans others, or otherwise violates the legal or social norms. Levin and Nolan (2007) define discipline problem as behavior that (1) intervenes with teaching, (2) intervenes with rights of other to learn, (3) is psychologically of physically unsafe, or (4) destroys property.

There are several conceptualizations of behavior and discipline in the school context but in this study the term classroom behavioral climate is used to describe behavior of students that might affect the behavioral climate of the classroom.

1.2 Classroom behavioral climate

Classroom behavioral climate (CBC) is a relatively new term in behavioral sciences. The term and a scale measuring CBC were developed in Finland as part of development and research of class-wide intervention on behavior (see Närhi, Kiiski, Peitso & Savolainen 2014; Närhi, Kiiski & Savolainen 2017).

CBC is based on Levin and Nolan's (2007) model of discipline problems. CBC is based on the following four components: (1) students' possibilities to study and concentrating on teaching; (2) disruptive behavior; (3) physical and psychological safety; and (4) caring for the physical environment (Hoffman et al., 2018). Both the behavior of teacher and students constitute to classroom behavioral climate (Hoffman et al. 2018).

At the core of CBC is the idea of disruption of learning but the elements of physical and psychological safety and caring for the physical environment are added to it. Defining a discipline problem only as disruptive behavior that halters learning of students leaves out behaviors that do not necessarily interfere with teaching or learning activities but have effects on the learning environment (Levin and Nolan, 2007). There are behaviors that can affect the physical or psychological safety of the learning environment, which is one of the key components of a positive learning climate (Kutsyuruba, Klinger & Hussain, 2015; Thapa et al., 2013). Caring for the physical environment describes behavior that affects the physical learning environment, which is also an element of wider learning climate (Kutsyuruba, Klinger & Hussain, 2015; Thapa et al., 2013).

1.3 Construct validity

Validity is a complex and multidimensional aspect of test development. Traditionally, there has been a view of three types of validity of (1) content validity, (2) criterion-related validity, and (3) construct validity (Messick, 1995). The Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2010) are not using these distinctions between different types of validity anymore but different kind of sources of validity evidence. Validity is seen as unitary concept and defined as the degree to which evidence and theory support the interpretations of test scores for proposed uses of test (AERA, APA, & NCME, 2010). This contemporary model of validity is also referred to as the unified construct-based model of validity (Dimitrov, 2010).

Messick (1995) specifies six aspects of the unified model of construct validity: (1) Content aspect, including evidence of content relevance, representativeness, and technical quality; (2) Substantive aspect, referring to theoretical rationales for the interpretations of test responses together with empirical evidence supporting the theoretical manifestation in practice; (3) structural aspect, examining the fidelity of the construct structure; (4) generalizability aspect, meaning the extent to which score properties and interpretations generalize to and across different populations and contexts; (5) external aspect, requiring convergent and discriminant evidence of convergent of how a measure relates to other measures, as well as evidence of criterion relevance and applied utility; (6) consequential aspect, examining the implications of score interpretations as well as the actual and potential consequences of test use with regard to issues of bias, fairness, and justice. Validation process involves gathering different validity evidence for a sound scientific basis for the proposed score interpretations (AERA, APA, & NCME, 2010).

In this study the structural and generalizability aspects of the CBC scale are examined. Firstly, analyses of the internal structure of the CBC scale are made to examine the degree to which the hypothesized factor structure of the CBC fits the empirical data. Secondly, it is examined how well does theory-based structure of

CBC generalize between Finnish and Greek primary school contexts. If comparisons between different groups on a construct are to be made, invariance of the measurement properties of the construct need to be ensured for valid comparisons (Dimitrov, 2010; Brown, 2006).

1.4 Measurement invariance

Measurement invariance assesses the equivalence of a construct's measurement properties across groups or time and demonstrates that a construct has the same meaning to those groups or across repeated measurements (Putnick & Bornstein, 2016). Measurement invariance is essential because it is a prerequisite to comparing group means (Putnick & Bornstein, 2016, Dimitrov, 2010). Vanderberg and Lance (2000) compiled a summary of recommended practices on examining measurement invariance and they separate two aspects of invariance: (1) measurement invariance, concerning testing of relationships between measured variables and latent constructs, and (2) structural invariance, examining latent variables themselves. According to Dimitrov (2010), measurement invariance is part of factorical invariance, there being three aspects of factorial invariance: configural invariance, measurement invariance, and structural invariance. Configural invariance is either seen as part of measurement invariance or a prerequisite to it.

Commonly considered measurement invariance steps are: (1) configural invariance, equivalence of model form or pattern; (2) metric invariance, equivalence of factor loadings; (3) scalar invariance, equivalence of item intercepts or thresholds; and (4) residual invariance, equivalence of item residuals or unique variances (Putnick & Bornstein, 2016; Vandenberg & Lance, 2000). Measurement invariance is assessed on three levels: weak, strong and strict). Weak measurement invariance requires that both configural and metric invariance are established. Under weak measurement invariance the relations between the latent factor and external variables can be compared across groups (Dimitrov, 2010). Strong measurement invariance requires also scalar invariance. Under strong measurement invariance, the comparison of factor means across groups is permissible (Dimitrov, 2010). For strict measurement invariance, in addition of earlier levels of invariance, invariance of residual must be in place. Strict invariance provides evidence that the items are measured with the same precision in each group and the group differences on any item are due only to group differences on the common factors (Dimitrov, 2010). Meaning, that latent construct is measured identically across groups (Schoot, Lygtig & Hox, 2012).

Metric invariance, Structural invariance indicates invariance of factor variances and covariances. This is required only if variability of target constructs and correlational relationships among them are deemed relevant to the generalizability aspect of validity. (Dimitrov, 2010.)

1.5 Research questions

The purpose of this study is to examine factorical invariance, specifically the configural and metric invariance, of the 17-item teacher classroom behavioral climate (CBC) scale in Finland and Greece with confirmatory factor models. Following three research questions were formed for this purpose:

1. Does the theorized correlating four-factor model of the CBC fit the samples of Finland and Greece?

2. Is the correlating four-factor baseline model of CBC configurally invariant?

2. What is the metric invariance of the CBC scale across the two samples?

2 RESEARCH METHODS

2.1 Research context

In this study validity of the CBC scale is looked at the context of Finnish and Greek education systems. Cultures and education system naturally differ between the countries but there are also similarities between them.

The Finnish education system is decentralized, and local autonomy of municipalities and teachers is high. On the contrary the Greek educational system is a highly centralized system. In Finland the Ministry of Education and Culture is the highest authority and is responsible for all publicly funded education in Finland. The Ministry is responsible for preparing educational legislation, all necessary decisions, and its share of the state budget for the Government. The Finnish National Agency for Education (EDUFI) operates under the Ministry. EDUFI is the national development agency responsible for early childhood education and care, pre-primary, basic, general, and vocational upper secondary education as well as for adult education and training. EDUFI prepares the national curriculum for primary education. National curriculum states the overall goals of education. Municipalities in Finland are in charge of financing and administration of education, and also staff appointments, together with schools themselves. Municipalities and even schools have their local curriculums, which detail how they implement the national curriculum.

In Greece the Ministry of Education and Religions is the main authority that makes all the policymaking and administration decisions related with financial issues, teaching staff appointment and curricula contents (SWPBS Erasmus, 2022). The Ministry of Education supervises the operation of pre-primary and primary schools. At an administrative lever, the schools of the whole country are divided in 12 geographical educational districts, which are supervised by a local authority working under the Ministry. (SWPBS Erasmus, 2022.)

Education systems overall have similar structure between the countries. Compulsory education in Finland applies to all 6–18-year-olds. However, before

2021 compulsory education applied only until 16 years of age. Compulsory education includes pre-primary, basic and, after the extension of the upper age limit, the upper secondary education. Basic education in Finland consists of primary (ages 7-12) and secondary education (ages 13-16). In contrast, compulsory education in Greece applies to all 4-15-year-olds, including pre-primary (ages 4-5), primary (ages 6-12) and secondary education (ages 13-15) (Eurydice, 2023). So in Greece compulsory educations starts earlier but in Finland ends later.

PISA is the OECD's Program for International Student Assessment, which gives insight on different countries' education systems. In PISA 2018 study on students' academic performance Finland was statistically significantly above the OECD average in all three categories, reading, mathematics and science, but Greece was below the OECD average in all of them (OECD, 2019a, 57-61). However, behavior problems are prevalent issue in both countries - in PISA 2018 schools' disciplinary climate was significantly below OECD average in both countries (OECD, 2019b, 68). It is not surprising, that in both countries research and development of behavior support in schools has gained interest.

Both Finnish and Greek partners participated in the Erasmus SWPBS project or the "Building School-Wide Inclusive, Positive and Equitable Learning Environments Through A Systems-Change Approach" (SWPBS) ERASMUS+ Key Action 3 Policy Experimentation program. SWPBS Erasmus was a research project between 2019 and 2022 funded with the support from the European Commission which aimed to establish an inclusive non-discriminatory social culture and necessary socio-emotional and behavioral supports for all children in a school across Cyprus, Finland, Greece, Romania (see www.pbiseurope.org). University of Jyväskylä was the research partner of the project from Finland. Dieythinsi Protovathmias and Deyterovathmias Ekpaideysis Thessalonikis (KMAKEDPDE) was the local authority of the project in Greece working with Aristotle University of Thessaloniki (AUTH) (SWPBS Erasmus, 2022).

As part of the project, School Wide Positive Behavior Supports (SWPBS) based intervention on behavior modelled after the Finnish ProKoulu program was implemented in Cyprus, Greece, and Romania. ProKoulu was a Finnish research project between 2013 and 2016 where SWPBS based model on school's behavioral climate and students' and teachers' wellbeing was examined (see www.prokoulu.fi). The project was organized together with University of Eastern Finland, University of Jyväskylä and Niilo Mäki institute, a Finnish research and development center focusing on learning and learning difficulties, funded by the Finnish Ministry of Education and Culture.

2.2 Data and participants

The data of this study consists of teacher evaluations of classroom behavior climate (CBC) from the baseline measurements from ProKoulu and Erasmus SWPBS projects, collected during Autumn semesters 2013 and 2019 respectively.

The initial measurement of the ProKoulu project in Finland targeted all primary education teaching staff (N= 1386) and students at school years 2-6 of 70 schools across Eastern Finland. Data of this study consists of 694 teacher evaluations of CBC from 70 schools.

In the SWPBS Erasmus project in Greece 270 teachers participated in the initial data collection phase from 30 schools across the prefectures of Thessaloniki and Halkidiki and Imathia (SWPBS Erasmus, 2022). The 30 schools had 472 teachers in total. One participant did not have data on CBC so Greek data used in this study consist of 269 teacher evaluations of CBC.

The participants in both projects were in-service teachers working in primary schools. In Finland some cojoined primary and secondary schools participated in the study, so some teachers worked in primary and secondary education. All teachers of the participating schools were targeted in both studies, meaning the sample consists of variety of class teachers, special education teachers and subject teachers. Data descriptives are seen in table 1.

Table 1 *Descriptives of Finnish and Greek data*

	Finland	Greece
N	694	269
Gender		
women	520 (75%)	197 (73%)
men	174 (25%)	72 (27%)
Age (years)		
range	23-63	21-66
M	44	49
SD	9,2	8,0
Work experience (years)		
range	0-38	3-38
M	18	21
SD	9,0	7,2
Experience in this school (years)		
range	0-34	1-32
M	10	9
SD	8,4	7,1

2.3 Data Collection

Teacher classroom behavioral climate (CBC) scale was developed originally for Finnish behavior intervention research. As part of the SWPBS Erasmus project, the Finnish scale was translated to English, and Greek and Romanian translations were made from English language. In the SWPBS Erasmus (2022) project the European Social Survey (ESS) translation guidelines were implemented as part of the TRAPD procedures. TRAPD is acronym for Translation, Review, Adjudication, Pre-testing and Documentation, the five integral procedures for ESS translation and assessment (SWPBS Erasmus, 2022). Data of this study consist of answers to Finnish and Greek language scales.

In the ProKoulu study data was collected using online questionnaires. The links to the questionnaires were sent to the teachers based on lists provided by the school administrations. In the SWPSB Erasmus project's baseline measurement teachers filled in paper questionnaires on site at the schools. In the project, a team of external coaches was compiled to oversaw training and coaching of school staff. External coaches were also in charge of data collection at schools accompanied by school counselors from the local authority KMAKEDPDE.

The teacher CBC scale consists of 17 items. Teachers were asked to answer how well each statement describes the working conditions of their classroom during instruction on 6-step Likert-scale (not at all – to a great deal). This means that the data is ordered categorical or ordinal.

The scale has four sub-scales: (1) students' possibilities to study and concentrating on teaching (4 items, for example "Students work peacefully on assignments during instruction."); (2) disruptive behavior (5 items, for example "There's inappropriate movement during instruction."); (3) physical and psychological safety (5 items, for example "Students feel comfortable answering even when they are unsure of the right answer." or "Students hit or threat to hit each other."); and (4) caring for the physical environment (3 items, for example "Students use classroom equipment appropriately.").

Items 4,5,6,7,8,9,10,11,13 and 16 were reversed, so that in all variables a higher value indicates more positive behavioral climate. Means, standard deviations and reliabilities of the sub-scales and the whole CBC scale are seen in table 2 below.

Table 2 ***Means, standard deviations and reliabilities (Cronbach's Alpha) of teacher classroom behavioral climate (CBC) scale and its sub-scales***

	Finland			Greece		
	M	SD	α	M	SD	α
students' possibilities to study and concentrating on teaching	3.89	.75	.82	4.57	.78	.81
disruptive behavior	3.63	.93	.89	4.15	.86	.87
physical and psychological safety	4.59	.78	.74	4.71	.71	.70
caring for the physical environment	4.18	.77	.66	4.52	.79	.65
CBC	4.07	.68	.91	4.33	.75	.91

Additionally, for multigroup analysis to work on ordinal data, some variables in the Finnish data had to be recoded. In the Greek data variables 1,2,5,8,10 ja 16 didn't have responses to all Likert-scale values. Variables were missing responses to value 1, except variable 16, which had missing responses to value 2 instead. This means that these variables in the Greek data had 5 categories instead of the intended 6. These variables were recoded so that values 1-2 were combined into a single category, so that both countries' data had the same number of categories in corresponding variables. Correlation table of all the scale variables used in the analyses are in Appendix 1.

2.4 Data Analysis

In this study the structural validity of the CBC scale was assessed with confirmatory factor analysis (CFA). CFA provides a comprehensive evaluation of the internal structure of a rating scale by testing the fit of the hypothesized factor structure to observed data (Brown, 2006). Analyses were performed using the Mplus Statistical Package Version 8 (Muthén & Muthén, 1998–2017).

Since the items are supposed to measure one construct, classroom behavioral climate, having four compounds it was expected that the items load to four factors according to the theory-based scale structure. This dimensionality was examined by CFA.

First, it was examined whether the theorized four-factor model fits in both groups, and baseline or best-fitting model for each group was identified. Two models where all parameter estimates were allowed to vary across the Finnish and Greek samples were tested and compared: a one-factor model where all items are loading to one factor and a four-factor model, where items loaded to the factors according to the theory-based scale structure.

Weighted least square mean and variance adjusted (WLSMV) estimation was used because of ordinal data. WLSMV is the recommended estimator in Mplus when using categorical data (Muthén & Muthén, 1998–2017). WLSMV is a robust estimator which does not assume variables to be normally distributed and is the most optimal option for modelling categorical or ordered data (Brown, 2006).

For evaluating CFA model fit, Mplus provides several indicators of goodness-of-fit. In this study goodness-of-fit was determined based on following fit indices: chi square (χ2), the root mean square of the approximation (RMSEA), comparative fit index (CFI), Tucker-Lewis index (TLI), and standardized root mean square residual (SRMR). Goodness of fit indices can be classified into absolute and incremental fit indices (Chen, 2007; Brown, 2006). Absolute fit indices assess the degree to which the model-implied covariance matrix matches the observed covariance matrix and in contrast, incremental fit indices assess the degree to which the tested model is superior to an alternative model in reproducing the observed covariance matrix (Chen, 2007). $\chi2$, RMSEA and SRMR are absolute fit indices, and CFI and TLI are incremental. For absolute indexes smaller number indicates a better model fit and for incremental indexes a larger number indicates a better fit.

A nonsignificant $\chi2$ test indicates a good model fit (Brown, 2006). However, $\chi2$ test is sensitive to the sample size, meaning that it can reject adequate models if the sample is large, and when the sample is small it can fail to reject poor models (Schoot, Lugtig & Hox, 2012). Additionally, structural equation models of reasonable complexity essentially never fit real data based on the test of absolute fit like the chi-square test (Savalei, 2021).

Generally, for RMSEA values less than 0.05 are considered good, values between 0.06-0.08 adequate and values over 0.1 poor. For CFI and TLI, that values above 0.95 represent an excellent model fit and 0.90–0.95 represent reasonable fit. Lastly, SRMR values below 0.05 are considered good and below 0.08 acceptable. (Brown, 2006.)

When the group-specific baseline models were first identified, the models were entered into a multiple group analysis in to test for configural invariance. Lastly, factor loadings were set to be invariant across the country models to investigate the metric invariance of the scale. The configural model (freely loading factors across measurements) was compared to a metric model (factor loadings fixed equally across the two countries).

The Satorra-Bentler (2001) scaled $\chi2$ test was used to compare the more restricted (metric) model with the less restricted (configural) model. Statistically significant $\chi2$ test means that loadings have statistically significant difference between the models and loadings are not invariant. Because WLSMV estimation was used, difference testing was executed with DIFFTEST option in Mplus (Muthén & Muthén, 1998-2017). However, since the difference test has the same problems concerning sample size and model complexity as the general $\chi2$ test, model invariance was also examined based on modification indexes. Chen (2007) suggests that a change of RMSEA and SRMR less than 0.015 and change of CFI less than 0.01 between models indicates invariance between models.

2.5 Ethical Solutions

Responsible Conduct of Research (RCR) was followed during the study. Guidelines for the responsible conduct of research have been published by the Finnish National Board on Research Integrity (TENK), which is appointed by the Ministry of Education and Culture in Finland. The principles of integrity, meticulousness, and accuracy in conducting research, and in recording, presenting, and evaluating the research results endorsed by the research community (TENK, 2012) have been followed in this study.

Data used in this study was collected beforehand, so, ethical questions concerning data acquisition are limited to accessing the data. Agreement about the use of research data in master's thesis was made in the beginning of master's thesis process with the representative of the Board of Principal Investigators from ProKoulu and SWPBS Erasmus projects. This agreement includes researchers' rights, responsibilities, and obligations, principles concerning authorship, and questions concerning archiving and accessing the data during the master's thesis process and after it.

However, in both projects the data in this study was attained from, the necessary research permits had been acquired and the preliminary ethical reviews were conducted. In the ProKoulu project researchers applied and got an ethical review from the University of Eastern Finland (UEF) Committee on Research Ethics prior to the commencement of the research. Also, local municipalities' policies were followed when gathering research permits.

In SWPBS Erasmus project in all partner countries ethical reviews were applied from local ethics committees. Greek researchers applied and got the review from the Research Ethics Committee (REC) of the Aristotle University of Thessaloniki. Research permits from schools were gathered in collaboration with local authorities, which is Greece was KMAKEDPDE.

All teachers, students and parents of students from the schools that participated in the project were given information on the purpose of the study. Participation to the study was voluntary and informed consent was given by the participants. A confidentiality statement was included with informed consent explaining research team's compliance to personal data protection laws.

Data used in this study does not include any personal or sensitive information. In the ProKoulu and SWPBS Erasmus projects anonymity of participants identities has been ensured by pseudonymization. Data storing and processing has been done carefully following the data privacy guidelines of University of Jyväskylä.

Lastly, in carrying out this study and publishing its results, the work and achievements of other researchers are given proper credit by respecting their work and citing appropriately.

3 RESULTS

3.1 Factor structure of the classroom behavioral climate scale

The one-factor model showed a poor fit with the data in both countries. Finland: $\chi2$ (119) = 2075.43, p< .001, RMSEA = .15, SRMR = .08, CFI = .90, TLI = .88. Greece: $\chi2$ (119) = 1085.44, p < .001, RMSEA = .17, SRMR = .09, CFI = .84, TLI = .82

In Finland the correlating four-factor model showed a good fit on three indices: $\chi2$ (113) = 1000.13, p < .001, RMSEA = .11, SRMR = .05, CFI = .95 TLI = .94. In Greece the correlating four factor model showed a more adequate fit but not a good level: $\chi2$ (113) = 692.95, p < .001, RMSEA = .14, SRMR = .08, CFI = .90, TLI = .89.

Modifications were made to increase the model fit by freeing error covariances between some items. In the Finnish model one error covariance between two items was freed: item 15 ("Students use classroom equipment appropriately.") and item 17 ("Students leave classroom tidy before they go home."), *cov*(15,17) = .56. Both items are part of the *caring for the physical environment* factor. Modified model showed good fit on the same three indices, SRMR, CFI and TLI, also having a lower $\chi2$ and RMSEA <0.1: $\chi2$ (112) = 763.25, p < .001, RMSEA = .09, SRMR = .05, CFI = .97, TLI = .96.

In the Greek model two error variances between items were freed. First, between item 10 ("Students call each other names.") and item 13 ("Students hit or threat to hit each other.") *cov*(10,13)= .65. Both items are part of the *physical and emotional safety* factor. Secondly, between item 14 ("Students respect each other's personal space.") with item 15 ("Students use classroom equipment appropriately."), *cov*(14,15) = .59. These items are on different factors, on *physical and emotional safety* and *caring for the physical environment* factors respectively. The modified model showed acceptable fit on three indices, SRMR, CFI and TLI: $\chi2$ (111) =531.67, p < .001, RMSEA = .12, SRMR = .07, CFI = .93, TLI = .92.

Standardized factor loadings for the modified four factor models are presented in table 3 and factor intercorrelations in table 4.

Table 3. Standardized factor loadings for the 17-item teacher CBC scale (modified correlated four-factor model)

Factor	Item number[1]	Finland	Greece
Studying and concentrating on teaching	1	.91	.83
	2	.86	.85
	3	.87	.86
	4	.69	.69
Disruptive behavior	5	.87	.85
	6	.78	.84
	7	.78	.79
	8	.90	.84
	9	.71	.71
Physical and emotional safety	10	.85	.70
	11	.81	.49
	12	.35	.63
	13	.80	.52
	14	.67	.67
Caring for the physical environment	15	.58	.78
	16	.79	.72
	17	.50	.75

Table 4. Factor intercorrelations (modified correlated four-factor model): Finland in lower diagonal, Greece in the upper diagonal.

	1	2	3	4
Studying and concentrating on teaching	-	.76***	.83***	.74***
Disruptive behavior	.84***	-	.77***	.68***
Physical and emotional safety	.66***	.70***	-	.94***
Caring for the physical environment	.79***	.75***	.88***	-

***p<.001

3.2 Configural invariance

Multiple group analysis showed adequate fit on three indices, SRMR, CFI and TLI, showing reasonable evidence on configural invariance of the theorized correlating four factor model of CBC: $\chi 2$ (285) =2190.23, p < .001, RMSEA = .12, SRMR = .06, CFI = .92, TLI = .93.

3.3 Metric invariance

The fit indices of the metric model showed adequate fit on three indices, SRMR, CFI and TLI: $\chi 2$ (298) =2272.325, p < .001, RMSEA = .12, SRMR = .06, CFI = .92, TLI =.93. The Satorra-Bentler scaled difference chi-square test was significant, $\Delta\chi 2(13)$ = 176,97, p <.001. This indicates statistically significant variance between factor loadings between the countries. However, the change of model fit indexes indicates invariance between the models. ΔSRMR = .004, which is less than .015 and ΔCFI=.003 which is less than .01, suggesting invariance of factor loadings. The change in RMSEA not evaluated because the index was over recommended thresholds; still two of three primary indices reported change well below adequate ranges.

4 DISCUSSION

Examining validity in different populations is important because the validity of a measurement tool may not generalize across different groups of people. The validation process can involve adapting the measurement tool to be culturally sensitive or gathering validation evidence to see if the same construct is found in different populations. The purpose of this study was to examine configural and metric invariance of the classroom behavioral climate (CBC) scale across Finnish and Greek primary school contexts as part of construct validation of the scale. The measurement equivalence of CBC scale between different populations had not been examined before this study. The results showed support for both configural and metric invariance of the scale.

Examining measurement validity is crucial to ensure that the results obtained from a measurement tool are reliable and meaningful, and that any decisions or actions based on those results are appropriate and accurate. Measurement invariance is a prerequisite for comparing different populations on a construct. If the measures relied on do not have the same meanings across different groups, the conclusions drawn from a study may be invalid or biased (Chen, 2007).

The CBC scale is a self-reporting scale, which are common in social and behavioral sciences, but such measures can suffer from various measurement errors. These errors are associated heavily with generalizability, validity across different contexts and populations. In previous research, it has been discovered that cultural differences affect the process and results of user research (Lee et al., 2007). Research also shows response bias, such as social desirability, acquiescence and extreme response choice, being associated with socio-economic development of countries (OECD, 2019c).

Overall, invariance testing of behavioral phenomena at school context has not been widely examined. School climate is a widely studied phenomena, that is linked with classroom behavioral climate. Measurement invariance of school climate between different populations withing countries, for example, gender, ag and race, has been reported (e.g., La Salle et al., 2021; Waasdorp et al., 2020). Examination between different countries is not so common, but for example Shukla and colleagues (2019) established partial invariance of school climate dimension between United States and Mexico.

In PISA study however, for each item and scale, analyses on the invariance of item parameters across countries and languages within a country were conducted (OECD 2019c). In addition, in PISA study similar and high reliability values across countries are seen as a good indication of having measured reliably across countries (OECD, 2019c). In this study reliabilities of CBC scale and subscales were on a good level and very similar between Finnish and Greek samples, seen in table 2. In PISA 2018 both student and teacher disciplinary climate indexes had similar reliabilities and high level of invariance between countries (OECD, 2019c, table 16.4, table 16.128). The teacher index however had more countries with unique parameters (OECD, 2019c, table 16.128).

Results of this study showed that the theorized four-factor structure fit well in the Finnish data and adequately in the Greek data and support for configural invariance was established. Metric invariance was supported by the minimal change in SRMR and CFI values. Measurement invariance is essential because it is a prerequisite to validly comparing group means. However, strong measurement invariance, meaning all configural, metric and scalar invariance, needs to be established for group mean comparison. In this study evidence for only weak measurement invariance was established. This means that comparison across groups on relations between the latent factor and external variables can be made (Dimitrov, 2010). Nonetheless, these results give a positive outlook on the theorized four factor model of CBC and the use of the scale in Finnish and Greek primary school contexts. However, further examination of the measurement invariance of the scale is needed and these results should be interpreted with caution due to limitations of the models and the data.

First, baseline models for the theorized four factor structure of CBC were established in both countries separately. The results indicate that the theory based four factor structure of the construct of CBC is found in both countries.

However, it is important to notice, that *physical and emotional safety* and *caring for environment* factors have high in-between correlation in both samples, especially in Greece. Uniqueness of safety and caring factors in Greece needs to be considered with caution, since most of the variance is shared by the two factors.

In addition, error covariances between some items were freed in both baseline models. Freeing error covariances based on modification indices is common but not unproblematic. There might be theoretical or substantial explanations for the correlated residuals, but the underlying cause cannot be known, only speculated. When model is changed based on modification indices theoretically driven process becomes data driven process (Landis et al., 2009). In this study, there are substantial similarities between the variables, that could explain the correlated residuals. For the further validation of the scale, the Finnish and Greek language scales should be examined to determine if there are possible explanations of the correlated residuals in the local translations of the scales. Items are often deleted based on significantly correlated residuals, but in this study the invariance of the whole scale structure was in focus, so no deletions were made.

There are also some other indicators that removing some items from the country models could be justified. In the Finnish model item 12 has a loading of 0.35 (see table 2) which is significantly lower than all the other variable loadings. Loadings below 0.4 are commonly considered inadequate (e.g., Whitley & Kite, 2018, Brown, 2006). Correlations of the item 12 are also relatively low, seen in Annex 1, but they are all statistically significant. In contrast, in Greek data item 11 does not have significant correlations with all items of the scale and of the safety factor it belongs in. Interestingly, both item 11 (""Students make fun of classmates, who give wrong answers.") and item 12 ("Students feel comfortable answering even when they are unsure of the right answer.") are part of the safety sub-scale and relate to students answering to questions during instruction. Examination of translations of these items could be beneficial but these behaviors might also be harder to evaluate from teachers' perspective compared to other items in the safety factor.

When considering at all the models in this study, generally three out of five fit indices were on a good level. It was expected that $\chi 2$ test would reject the models based on the problems concerning sample size, complexity of factor model and ordinal data (Schoot, Lugtig & Hox, 2012; Savalei, 2021). The commonly used cutoff values for fit indices are based on Hu and Bentler's (1999) findings. However, use of these cutoff values has been criticized (e.g., Marsh, Hau, & Wen, 2004). It is important to remember that these cutoff values are general rule of thumb and do not necessary generalize to complex models, different sample sizes or non-normality of data.

There are also limitations concerning the data in this study. The data in this study was ordinal, which sets some limitations concerning the analyses. However, research has showed concerns about the problematic performance of categorical fit indices (Savalei, 2021; Xia & Yang, 2018, 2019). For example, RMSEA and CFI have been found to be sensitive on variables such as the estimator used, the number of categories in the data, and the values of the thresholds (Xia & Yang, 2018, 2019). Analyses could have alternatively been made treating the data as continuous, but treating ordered categorical data as continuous often violates the assumption of multivariate normality, also distorting the factor structure across groups and potentially producing inaccurate results (Lubke & Muthén, 2004).

The data was also skewed, which resulted also in recoding of some variable values. Recording influences item variance. Most of the discipline problems at school are mild (Skiba et al., 1997), and several skewed variables measured more severe behavior, for example item 16 "Students deliberately break the classroom equipment." so it makes sense that few teachers reported behaviors like that being common. Also, general statements like item 1 "There is peaceful working climate during instruction." or item 5 "It's too noisy during lessons" were skewed, which reflects classroom behavioral climate being on a good level in the data on average, as seen on table 2. Mean of CBC in Finland was 4.07 and in Greece 4.33 on scale from 1 to 6.

Considering the participants in this study, in both countries participants were from a restricted area of the country. Finnish ProKoulu data collected from Eastern Finland and Greek SWPBS Erasmus data was collected from one educational district, so wider examination of CBC in both countries could be done. In this study information about participant occupation or what school years teachers work in was not available. Measurement invariance between different teacher occupations or between for example primary and secondary schools could be interesting research topic for the future.

In this study teacher CBC scale was examined, but also the factor structure and measurement invariance of the student CBC scale would be worth examining also. It would be interesting to compare the factor structure of teacher and student scales. Previous research has noted that teacher and students' assessment of classroom behavioral climate differ (Holopainen et al., 2009), but are the elements of CBC the same for both teachers and students?

Overall, structural validity of the CBC scale needs further examination. Since configural and metric invariance of the CBC scale across Finland and Greece was established in this study, logical next step would be examining scalar and other levels of invariance between the countries. Also, since the scale has been used in other countries, invariance across them should be assessed also. For valid cross-cultural examinations to be made, it needs to be established whether the same construct is found in different contexts.

This study is a good beginning on validating CBC scale in different populations. If the CBC scale properties are proven equivalent between different populations, further comparison and examinations on classroom behavioral climate in different contexts can be made. Information regarding CBC can be utilized in development of education systems and effective behavior support methods, which is valuable since behavior problems are a major issue in school contexts around the world.

References

Algozzine, B., Wang, C., & Violette, A. S. (2011). Reexamining the Relationship Between Academic Achievement and Social Behavior. Journal of positive behavior interventions, 13(1), 3-16. https://doi.org/10.1177/1098300709359084

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (AERA, APA, & NCME). (2020). The Standards for Educational and Psychological Testing. American Educational Research Association.

Bollmer, J. M., Milich, R., Harris, M. J., & Maras, M. A. (2005). A Friend in Need: The Role of Friendship Quality as a Protective Factor in Peer Victimization and Bullying. Journal of interpersonal violence, 20(6), 701-712. https://doi.org/10.1177/0886260504272897

Bowen, N. K., & Masa, R. D. (2015). Conducting Measurement Invariance Tests with Ordinal Data: A Guide for Social Work Researchers. *Journal of the Society for Social Work and Research, 6*(2), 229. https://doi.org/10.1086/681607

Brown, T. A. (2006). Confirmatory factor analysis for applied research. Guilford Press.

Boyle, G. J., Borg, M. G., Falzon, J. M. & Baglioni, A. J. (1995). A structural model of the dimensions of teacher stress, British Journal of Educational Psychology, 65(1), 49–67. https://doi.org/10.1111/j.2044-8279.1995.tb01130.x

Charles, C.M. (2005). Building classroom discipline. (8th ed.). Pearson.

Cheema, J. & Kitsantas, A. (2014). Influences of disciplinary classroom climate on high school student self-efficacy and mathematics achievement: a look at gender and racial-ethnic differences", International Journal of Science and Mathematics Education, Vol. 12/5, pp. 1261-1279, http://dx.doi.org/10.1007/s10763-013-9454-4.

Chen, F. F. (2007). Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. Structural equation modeling, 14(3), 464-504. https://doi.org/10.1080/10705510701301834

Dimitrov, D. M. (2010). Testing for Factorial Invariance in the Context of Construct Validation. Measurement and evaluation in counseling and development, 43(2), 121-149. https://doi.org/10.1177/0748175610373459

Eurydice, European Education and Culture Executive Agency. (19 April 2023). National Education systems, Greece, overview. From https://eurydice.eacea.ec.europa.eu/national-education-systems/greece/overview

Finn, J. D., Fish, R. M., & Scott, L. A. (2008). Educational Sequelae of High School Misbehavior. The Journal of educational research (Washington, D.C.), 101(5), 259-274. https://doi.org/10.3200/JOER.101.5.259-274

Hakanen, J., A. Bakker & W. Schaufeli. (2006). Burnout and work engagement among teachers. Journal of School Psychology, Vol. 43/6, pp. 495-513, http://dx.doi.org/10.1016/j.jsp.2005.11.001.

Hinshaw, S. P. (1992). Externalizing behavior problems and academic underachievement in childhood and adolescence: Causal relationships and underlying mechanisms. Psychological Bulletin, 111(1), 127–155. https://doi.org/10.1037/0033-2909.111.1.127

Hochweber, J., Hosenfeld, I., & Klieme, E. (2014). Classroom composition, classroom management, and the relationship between student attributes and grades. Journal of Educational Psychology, 106(1), 289–300. https://doi.org/10.1037/a0033829

Hoffmann, L., Närhi, V., Savolainen, H., & Schwab, S. (2021). Classroom behavioural climate in inclusive education – a study on secondary students' perceptions. Journal of research in special educational needs, 21(4), 312-322. https://doi.org/10.1111/1471-3802.12529

Holopainen, P., Järvinen, R., Kuusela, J. & Packalen, P. (2009) Työrauha tavaksi. Kohtaaminen, toimintakulttuuri ja pedagogiikka koulun arjessa (Helsinki, Opetushallitus).

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

Karakus, Mustafa C., David S. Salkever, Eric P. Slade, Nicholas Ialongo, and Elizabeth Stuart. 2012. "Implications of Middle School Behavior Problems for High School Graduation and Employment Outcomes of Young Adults: Estimation of a Recursive Model."Education Economics 20 (1): 33–52. https://doi.org/10.1080/09645292.2010.511816.

Klassen, R. M., & Chiu, M. M. (2010). Effects on Teachers' Self-Efficacy and Job Satisfaction: Teacher Gender, Years of Experience, and Job Stress. Journal of educational psychology, 102(3), 741-756. https://doi.org/10.1037/a0019237

Kutsyuruba, B., Klinger, D. A., & Hussain, A. (2015). Relationships among school climate, school safety, and student achievement and well-being: A review of the literature. *Review of education (Oxford), 3*(2), 103-135. https://doi.org/10.1002/rev3.3043

La Salle, T. P., McCoach, D. B., & Meyers, J. (2021). Examining Measurement Invariance and Perceptions of School Climate Across Gender and Race and Ethnicity. Journal of Psychoeducational Assessment, 39(7), 800–815. https://doi.org/10.1177/07342829211023717

Lee, J., Tran, TT., Lee, KP. (2007). Cultural Difference and Its Effects on User Research Methodologies. In: Aykin, N. (eds) Usability and Internationalization. HCI and Culture. UI-HCII 2007. Lecture Notes in Computer Science, vol 4559. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-73287-7_16

Levin, J. & Nolan J. F. (2007). Principles of classroom management: A professional decision-making model (5. painos). Boston: Pearson.

Lubke, G. H., & Muthén, B. O. (2004). Applying multi-group confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. Structural Equation Modeling, 11, 514–534. https://doi.org/10.1207/s15328007sem1104_2

Malinen, O., & Savolainen, H. (2016). The effect of perceived school climate and teacher efficacy in behavior management on job satisfaction and burnout: A longitudinal study. Teaching and teacher education, 60, 144-152. https://doi.org/10.1016/j.tate.2016.08.012

Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. Structural Equation Modeling, 11, 320–341. https://doi.org/10.1207/s15328007sem1103_2

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Moos, R. (1979), Evaluating Educational Environments, Jossey-Bass, San Francisco, CA. https://doi.org/10.2307/1981414

Muthén, L.K. and Muthén, B.O. (1998-2017). Mplus User's Guide. Eighth Edition. Los Angeles, CA: Muthén & Muthén

Nurmi, J. E. 2012. Students' characteristics and teacher–child relationships in instruction: A meta-analysis. Educational Research Review, 7(3), 177-197. https://doi.org/10.1016/j.edurev.2012.03.001

Närhi, V., Kiiski, T., Peitso, S. & Savolainen, H. (2014). Reducing Disruptive Behaviours and Improving Learning Climates with Class-wide Positive Behaviour Support in Middle Schools. European Journal of Special Needs Education, 30(2), 274–285. https://doi.org/10.1080/08856257.2014.986913

Närhi, V., Kiiski, T. & Savolainen, H. (2017). Reducing disruptive behaviours and improving classroom behavioural climate with class-wide positive behaviour support in middle schools. British Educational Research Journal, 43, 1186–1205. https://doi.org/10.1002/berj.3305

OECD (2019a) PISA 2018 Results (Volume I): What Students Know and Can Do https://doi.org/10.1787/5f07c754-en

OECD (2019b) PISA 2018 Results (Volume III) What School Life Means for Students' Lives https://doi.org/10.1787/acd78851-en

OECD (2019c). PISA 2018 technical report. Chapter 16: Scaling procedures and construct validation of context questionnaire data. OECD Publishing. From https://www.oecd.org/pisa/data/pisa2018technicalreport/

Oliver, R. M., Wehby, J. H., & Reschly, D. J. (2011). Teacher classroom management practices: Effects on disruptive or aggressive student behavior. Campbell Systematic Reviews, 7(4), 1-55. http://dx.doi.org/10.4073/csr.2011.4.

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental review*, *41*, 71-90. https://doi.org/10.1016/j.dr.2016.06.004

Landis, R. S., Edwards, B. D., & Cortina, J. M. (2009). On the practice of allowing correlated residuals among indicators in structural equation models. In C. E. Lance & R. J. Vandenberg (Eds.), Statistical and Methodological Myths and Urban Legends: Doctrine, Verity and Fable in Organizational and Social Sciences (pp. 193-215). New York: Routledge https://doi.org/10.4324/9780203867266-16

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. Psychometrika, 66(4), 507–514. https://doi.org/10.1007/BF02296192

Savalei, V. (2021). Improving Fit Indices in Structural Equation Modeling with Categorical Data. Multivariate behavioral research, 56(3), 390-407. https://doi.org/10.1080/00273171.2020.1717922

Schaeffer, C. M., Petras, H., Ialongo, N., Masyn, K. E., Hubbard, S., Poduska, J., & Kellam, S. (2006). A Comparison of Girls' and Boys' Aggressive-Disruptive Behavior Trajectories Across Elementary School: Prediction to Young Adult Antisocial Outcomes. *Journal of consulting and clinical psychology, 74*(3), 500-510. https://doi.org/10.1037/0022-006X.74.3.500

Schoot, van de, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. European Journal of Developmental Psychology, 9(4), 486–492. https://doi.org/10.1080/17405629.2012.686740

Skiba, R. J., Peterson, R. L., & Williams, T. (1997). Office Referrals and Suspension: Disciplinary Intervention in Middle Schools. Education & treatment of children, 20(3), 295–315

Spilt, J. L., & Koomen, H. M. (2009). Widening the View on Teacher-Child Relationships: Teachers' Narratives Concerning Disruptive Versus Nondisruptive Children. School psychology review, 38(1), 86-101. https://doi.org/10.1080/02796015.2009.12087851

Shukla, K. D., Waasdorp, T. E., Lindstrom Johnson, S., Orozco Solis, M. G., Nguyen, A. J., Rodríguez, C. C., & Bradshaw, C. P. (2019). Does School Climate Mean the Same Thing in the United States as in Mexico? A Focus on Measurement Invariance. Journal of psychoeducational assessment, 37(1), 55-68. https://doi.org/10.1177/0734282917731459

SWPBS Erasmus, The "Building School-Wide Inclusive, Positive and Equitable Learning Environments Through A Systems-Change Approach" (SWPBS). (2022). D3.1 Implementation report for SWPBS Tier 1, Greece. From https://pbiseurope.org/documents/outputs/SWPBS_ImplementatioReport-Tier1_GR.pdf

TENK, Finnish Advisory Board of Research Integrity. (2012). Responsible conduct of research and procedures for handling allegations of misconduct in Finland. Guidelines of the Finnish Advisory Board on Research Integrity 2012. https://tenk.fi/sites/tenk.fi/files/HTK_ohje_2012.pdf

Thapa, A., Cohen, J., Guffey, S., & Higgins-D'Alessandro, A. (2013). A Review of School Climate Research. Review of Educational Research, 83(3), 357-385. https://doi.org/10.3102/0034654313483907

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. Organizational Research Methods, 3(1), 4–69. https://doi.org/10.1177/109442810031002

Waasdorp, T., Johnson, S. Shukla, K. D. & Bradshaw, C. P. (2020). Measuring School Climate: Invariance across Middle and High School Students, Children & Schools, Volume 42, Issue 1, January 2020, Pages 53-62, https://doi.org/10.1093/cs/cdz026

Wagner, M., Kutash, K., Duchnowski, A. J., Epstein, M. H., & Sumi, W. C. (2005). The Children and Youth We Serve: A National Picture of the Characteristics of Students With Emotional Disturbances Receiving Special Education. Journal of emotional and behavioral disorders, 13(2), 79–96. https://doi.org/10.1177/10634266050130020201

Whitley, Jr., B.E., & Kite, M.E. (2018). Principles of Research in Behavioral Science: Fourth Edition (4th ed.). Routledge. https://doi.org/10.4324/9781315450087

Xia, Y., & Yang, Y. (2018). The influence of number of categories and threshold values on fit indices in structural equation modeling with ordered categorical data. Multivariate Behavioral Research, 53(5), 731–755. https://doi.org/10.1080//00273171.2018.1480346

Xia, Y., & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. Behavior Research Methods, 51(1), 409–428. https://doi.org/10.3758/s13428-018-1055-2

Appendix 1

Classroom behavioral climate scale Spearman's correlations: Finland in lower diagonal, Greece in upper diagonal.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**
1	.	.61**	.68**	.41**	.58**	.43**	.41**	.32	.33**	.35**	.33**	.33**	.16**	.47**	.42**	.27**	.44**
2	.68**	.	.64**	.52**	.36**	.35**	.52**	.35**	.31**	.26**	.40**	.38**	.12	.48**	.49**	.21**	.40**
3	.72**	.69**	.	.43**	.49**	.41**	.45**	.42**	.34**	.36**	.34**	.38**	.24**	.52**	.50**	.25**	.45**
4	.46**	.49**	.49**	.	.51**	.42**	.49**	.38**	.31**	.25**	.39**	.39**	.16*	.42**	.41**	.19**	.39**
5	.66**	.58**	.57**	.56**	.	.65**	.53**	.57**	.57**	.43**	.20**	.28**	.33**	.38**	.35**	.31**	.44**
6	.54**	.50**	.49**	.46**	.62**	.	.53**	.59**	.51**	.45**	.14*	.26**	.35**	.33**	.40**	.28**	.41**
7	.52**	.48**	.48**	.43**	.65**	.56**	.	.55**	.44**	.36**	.24**	.47**	.22**	.33**	.35**	.26**	.39**
8	.62**	.57**	.55**	.51**	.69**	.67**	.68**	.	.61**	.53**	.07	.25**	.40**	.30**	.33**	.28**	.36**
9	.46**	.44**	.42**	.42**	.58**	.54**	.53**	.62**	.	.51**	.08	.23**	.32**	.22**	.26**	.20**	.33**
10	.39**	.39**	.43**	.42**	.45**	.42**	.45**	.54**	.41**	.	.12	.34**	.61**	.26**	.28**	.28**	.29**
11	.35**	.38**	.38**	.40**	.43**	.37**	.38**	.44**	.37**	.59**	.	.45**	.04	.44**	.41**	.23**	.29**
12	.20**	.28**	.21**	.17**	.18**	.14**	.23**	.19**	.16**	.23**	.29**	.	.27**	.43**	.44**	.37**	.30**
13	.30**	.30**	.31**	.38**	.36**	.37**	.30**	.38**	.30**	.58**	.48**	.11**	.	.20**	.24**	.37**	.23**
14	.38**	.41**	.40**	.32**	.32**	.33**	.28**	.37**	.30**	.44**	.37**	.20**	.57**	.	.74**	.36**	.54**
15	.37**	.38**	.39**	.34**	.34**	.36**	.30**	.37**	.31**	.31**	.29**	.27*	.22**	.34**	.	.34**	.58**
16	.33**	.33**	.35**	.44**	.40**	.38**	.35**	.40**	.32**	.51**	.51**	.22**	.48**	.34**	.40**	.	.34*
17	.37**	.36**	.39**	.32**	.32**	.32**	.31**	.34**	.28**	.25**	.25**	.25**	.22**	.30**	.61**	.31**	.

** <.001 * <.01