Psychological Test and Assessment Modeling, Volume 60, 2018 (2), 223-240

Measurement of teachers' reactive, preventive and proactive classroom management skills by student ratings – Results from a two-level confirmatory factor analysis

Christian Spoden[1] *& Katharina Fricke*[2]

Abstract

Student ratings can be an effective method to obtain measures of instructional quality although the usage of student ratings involves methodical challenges, raising the need for an adequate psychometric approach. Classroom management skills are an important aspect of instructional quality and a key competence of a teacher. In the present paper a two-level confirmatory factor analysis based on a shared cluster construct model according to Stapleton, Yang and Hancock (2016) is utilized with respect to the measurement of classroom management skills by student ratings. The approach is applied for the investigation of the dimensional structure of classroom management skills, assessed by means of a newly developed questionnaire. The results of the analysis support a three-dimensional structure with reliable measures of reactive, proactive, and preventive components of the construct of interest. These results are discussed in terms of implications for the assessment of classroom management skills using student ratings.

Keywords: two-level confirmatory factor analysis, classroom management, student ratings, instructional quality

[1] *Correspondence concerning this article should be addressed to:* Christian Spoden, German Institute for Adult Education – Leibniz Centre for Lifelong Learning, Heinemannstraße 12-14, 53175 Bonn, Germany. email: spoden@die-bonn.de

[2] University of Münster

Aspects of instructional quality are intensively studied by researchers in different strands of educational research (e.g. Muijs, & Reynolds, 2011; Nilsen & Gustafsson, 2016; Seidel, & Shavelson, 2007). Student (self-report) ratings are a frequently used source of information to determine and quantify the effects of instructional quality measures on learning (e.g., Seidel, & Shavelson, 2007). Student ratings facilitate comparisons across teachers over a potentially long time period and provide condensed information from a substantial number of informants with long-lasting experience within the class (Wagner, Göllner, Helmke, Trautwein, & Lüdtke, 2013). Compared to expert ratings of video-based observations of teaching sessions, which are often considered as gold standard of instructional research, student ratings are easily and immediately obtained in a survey study (Lüdtke, Trautwein, Kunter, & Baumert, 2006). In contrast to the assessment of some other components of instructional quality, student ratings for the assessment of classroom management (e.g., Dollase, 2012; Kounin, 2006), as one of the most important aspects of instructional quality and part of several instructional quality models (e.g., Brophy, 2000; Helmke, 2009; Klieme, Pauli, & Reusser, 2009), were also found to be generalizable across classes and school subjects (Wagner et al., 2013), useful in terms of predicting student achievement (Fauth et al., 2014; Kunter et al., 2008; Wagner et al., 2016), quite consistent over time, and in moderate to high agreement with teacher ratings (Wagner et al., 2016). However, the assessment of classroom management skills, which refer to a heterogeneous set of actions of a teacher to maximize the learning time available for content-related activities without disruption (Brophy, 2000; Seidel & Shavelson, 2007), by means of student ratings is challenging from a statistical point of view. Student ratings are not only prone to measurement error. The ratings are also assessed on the individual student level (within-level) but analyzed on the cluster or classroom level (between-level), which underscores the need for a multilevel measurement model (e.g., Fauth et al., 2014; Wagner et al., 2013; 2016). Additionally, the assessment of classroom management often assumes a unitary construct, which ignores the variability of preventive, proactive, and reactive actions and strategies of a teacher to handle order problems in the classroom. The assessment of different classroom management skills requires multidimensional measurement models (e.g., Brown, 2006; Reckase, 2009) to provide subtest scores of different components or factors of the construct of interest. The two-level confirmatory factor analysis (TL-CFA), which is just recently available in the lavaan package (Rosseel, 2012) for the R software (R Core Team, 2017), offers a valuable statistical approach to estimate profiles of classroom management skills from multidimensional student ratings. In the present paper the TL-CFA based on a shared cluster construct approach (Stapleton et al., 2016) is applied to investigate whether reactive, proactive, and preventive behavioral patterns of teachers' classroom management skills in physics education (as an important aspect of instructional quality) can be psychometrically distinguished based on student ratings.

This paper is organized as follows. First, two different approaches to conceptualize cluster-level constructs assessed by student ratings are differentiated, corresponding measurement models are described, and reasons for the application of the shared cluster construct model to student ratings on instructional quality measures and classroom management are explicated. Second, application of the model is further characterized in terms of prerequisites of data clustering, model fit, and reliability estimation. Third, measurement of three components of classroom management skills is described as field of application. Fourth, the research questions related to investigating the dimensional structure of student ratings on classroom management skills in physics education are given. Fifth, results from the statistical analysis, supporting a multidimensional structure, are provided. Sixth and concluding, these results are discussed concerning consequences for the assessment of classroom management skills using student ratings.

Two-level confirmatory factor analysis of student ratings

A common way to inspect multidimensionality of student ratings and to account for the clustered data structure (students nested in classrooms) is given by a TL-CFA framework. There are at least two different ways to conceptualize cluster-level constructs in a TL-CFA, either referred to as *reflective* and *formative constructs* (Lüdtke, Marsh, Robitzsch, & Trautwein, 2011), *contextual* and *climate constructs* (Marsh et al., 2012) or as *shared* and *configural constructs* (Stapleton et al., 2016), which is also the notation followed here.

Configural constructs include cluster aggregates of individual constructs. They incorporate the average of the individual configural construct at the cluster level as an indicator of the composition of that construct in the classroom. Following an example given in Marsh et al. (2012) for the application of this approach, big-fish-little-pond describes the effect that both individual student achievement (positively) and the classroom-average achievement (negatively) affect the academic self-concept of students. Due to individual differences with respect to personal attitudes, a strong agreement between the ratings of different students is not necessarily expected (Marsh et al., 2012; Stapleton et al., 2016). Configural constructs are represented in a two-level factor model, which involves simultaneous estimation of two factor analyses: one that accounts for the structure of items on individual level, and one that accounts for the structure of items on classroom level (Wagner et al., 2013).

Shared cluster constructs, on the other hand, represent a cluster characteristic at the cluster level assessed by means of individual level ratings. This refers especially to teacher characteristics rated by their students, for example classroom management skills of the teacher. Consequently, the shared construct approach assumes that there is a certain level of agreement across the student ratings, given that these ratings refer to the same teacher (e.g., Marsh et al., 2012; Stapleton et al., 2016). Although shared cluster constructs may also be represented in the two-level factor model, the shared cluster construct model (Figure 1, right-hand side) is an alternative specification where the construct of interest – although being assessed on the individual level – is specified only at the cluster level. Following the notation in Stapleton et al. (2016), assume that $\boldsymbol{y}$ is a $L \times 1$ vector of ratings on L items of an instructional quality questionnaire. This rating is decomposed into a within- and a between-clusters component by

$$\boldsymbol{y} = \boldsymbol{\eta}_{\boldsymbol{W}} + \boldsymbol{\eta}_{\boldsymbol{B}}\ . \qquad (1)$$

where $\boldsymbol{\eta}_{\boldsymbol{W}}$ is a corresponding vector of cluster mean-centered deviations and $\boldsymbol{\eta}_{\boldsymbol{B}}$ is a corresponding vector of cluster means. In the measurement part of the model, $\boldsymbol{\eta}_{\boldsymbol{B}}$ is related to the cluster-level latent variable or factor ξ_B by

$$\boldsymbol{\eta}_{\boldsymbol{B}} = \boldsymbol{\lambda}_{\boldsymbol{B}}\xi_B + \boldsymbol{\varepsilon}_{\boldsymbol{B}} \,. \qquad (2)$$

where $\boldsymbol{\lambda}_{\boldsymbol{B}}$ is a vector of the cluster-level factor loadings, and $\boldsymbol{\varepsilon}_{\boldsymbol{B}}$ is a vector of pertinent error terms on cluster level. The shared cluster construct model is identified when a constraint is set on the between-level, for example by fixing the first item factor loading or the factor variance to one. This model part can be extended to a multidimensional case with a vector $\boldsymbol{\xi}_{\boldsymbol{B}}$ of K factors representing subtest measures of the construct of interest. No measurement model is specified on the individual level; the model includes covariance estimates between the individual student ratings. The two-level specification offers correctly computed model parameters, standard errors, and model fit indices (e.g., Hox, 2010; Raudenbush & Bryk, 2002), contrary to a single-level model specification. Figure 1 shows both the two-level factor model on the left-hand side, which includes the corresponding measurement model on the individual level (indicated by subscript W instead of B), and the shared cluster construct model on the right-hand side, which includes covariance estimates between item i and j, θ_{ij}, on the individual level.

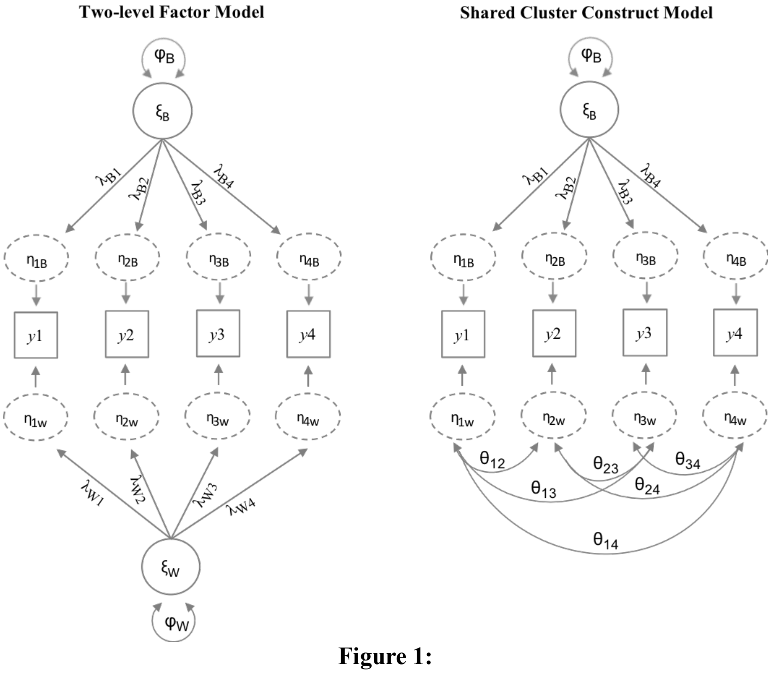

Figure 1:
Schematic illustration of a two-level factor model and a shared cluster construct model (adapted from Stapleton et al., 2016; error terms not shown for simplicity)

Both the two-level factor model and the shared cluster construct can be estimated by means of a maximum likelihood procedure (default estimator in the lavaan package). Maximum likelihood estimation with robust standard errors and a Satorra-Bentler scaled test statistic or with robust (Huber-White) standard errors and a scaled test statistic (asymptotically equal to the Yuan-Bentler test statistic) can be used in case of non-continuous response data (Rosseel, n.d.).

Degree of clustering, fit of the measurement model and reliability estimation

Although the two-level factor model and the shared cluster construct model can be compared empirically by means of information criteria, Stapleton et al. (2016, p. 494) argued that a ”[…]strong theoretical rationale should be provided if the model also includes any construct modeled at the within-cluster level[…]”. When the items were originally developed with the intention to solely assess skill sets or competences of teachers, this strong theoretical rationale is not always derived, while the shared cluster construct model is a straightforward statistical model to investigate the structure of classroom management skills of teachers assessed by means of student ratings.

To tests the adequacy of test data for the shared cluster construct model, Stapleton et al. (2016) suggested the following procedure:

1. Researchers provide the intraclass correlation coefficients ICC(1) and ICC(2) for each manifest variable.
2. Researchers provide evidence for the proposed measurement model (i.e. the relation between manifest and latent variables) by means of model fit statistics.
3. Researchers provide between-cluster composite reliability estimates of the measures.

The usage of intraclass correlation coefficients is well-established in multilevel research as a sufficient level of variability on both levels is a precondition for the application of multilevel models. The intraclass correlation ICC(1) (Shrout & Fleiss, 1979) refers to the degree of clustering of (manifest) variables and the agreement across the student ratings from the same classrooms, respectively. It is given by

$$\mathrm{ICC}(1) = \frac{\sigma^2_{\eta_B}}{\sigma^2_{\eta_B} + \sigma^2_{\eta_W}} \tag{3}$$

with $\sigma^2_{\eta_B}$ referring to the variability of the cluster level component and $\sigma^2_{\eta_W}$ referring to the variability of the individual level component. Values of ICC(1) > .05 are considered to be non-trivial (Geldhof, Preacher, & Zyphur, 2014), indicating suitability of a multilevel modeling approach. This coefficient is implemented in several software packages. The second coefficient ICC(2) (Raudenbush & Bryk, 2002) is considered as a measure of reliability of cluster components. It is given by

$$\mathrm{ICC}(2) = \frac{\sigma^2_{\eta_B}}{\sigma^2_{\eta_B} + \frac{\sigma^2_{\eta_W}}{n.}} \tag{4}$$

with $\sigma^2_{\eta_B}$ and $\sigma^2_{\eta_W}$ as defined before and $n.$ as the average cluster size. The coefficient ICC(2) can be utilized to determine the relevant number of single measures for reaching a given reliability level (Raudenbush & Bryk, 2002). Stapleton et al. (2016) argued that values of ICC(2) ≥ .50 indicate marginal degrees of reliability, values of ICC(2) ≥ .70 indicate acceptable degrees of reliability of the shared cluster construct.

Addressing the second criteria of providing evidence for the adequacy of the measurement model includes computation of absolute fit indices like CFI, TLI, RMSEA, and SMRM and referring to conventional cutoffs concerning these criteria (e.g., RMSEA ≈ .06, SRMR ≈ .08; Hu & Bentler, 1999) to evaluate the model fit. However, there is currently no general standard for absolute fit indices in TL-CFA, and in a recent study by Hsu, Kwok, Lin, and Acosta (2015) misspecifications on the cluster-level remained undetected by most common fit indices for structural equation models, except for the SRMR index computed only based on the cluster level (between-level SRMR, $\mathrm{SRMR_B}$). Additionally, relative fit indices like the information criteria AIC, BIC, and the sample-size adjusted BIC ($\mathrm{BIC_{adj}}$) can be used to compare two or more competing models.

While most applied research in social sciences relies on Cronbach's Alpha as reliability estimate, there is growing consensus that this coefficient is useful only under special conditions (in summary: Cho, 2016). Cho (2016) has given a summary of several composite reliability measures, which serve as an alternative to Cronbach's Alpha. He has also provided a Microsoft Excel Spreadsheet to compute the coefficients. The composite reliability measure for a CFA (correlated factors model in terms of Cho, 2016) is given by

$$\omega = \frac{(\sum_{i=1}^{k} \lambda_i)^2\ \varphi}{(\sum_{i=1}^{k} \lambda_i)^2\ \varphi + \sum_{i=1}^{k} \theta_{ii} + 2\sum_{i<j} \theta_{ij}} \tag{5}$$

where k is the number of items in a factor, φ refers to the factor variance, θ_{ii} refers to the residual variance (measurement error) of item i, and θ_{ij} refers to the covariance of measurement errors from items i and j. Note that ω is commonly estimated when φ is constrained to 1 to fix the scale. In case of the shared cluster construct model, Stapleton et al. (2016) argued to compute a between-level composite reliability measure based on the cluster-level factor loadings, the cluster-level factor variance, and the residual variances.

Application to the assessment of classroom management skills

Classroom management is an important teacher competence included in several instructional quality models (e.g., Brophy, 2000; Helmke, 2009; Klieme, Pauli, & Reusser, 2009) and often assessed by means of student ratings. Although no single definition of classroom management has been established, a general awareness of all student activities in the classroom, the organization of learning time and materials, the system of rules and rituals established, and the disciplinary actions and strategies applied by a teacher to handle order problems are among the skills commonly attributed to classroom management (e.g., Dollase, 2012; Kounin, 2006). The importance of teachers' classroom management skills for their students' academic achievement originates from a higher amount of learning time available for content-related activities without disruptions (Brophy, 2000; Seidel & Shavelson, 2007). Classroom management is based on the teachers' knowledge about the effectiveness of instructional activities, on the teachers' pedagogical and personal attitudes, and it also requires intensive verbal or non-verbal communication with their students (e.g., Doyle, 2011; Kounin, 2006).

A common way to distinguish different classroom management skills in preparation of classroom management profiles is given by a classification in reactive, preventive, and proactive components of classroom management (Borich, 2007; Fricke, van Ackeren, Kauertz, & Fischer, 2012; Helmke, 2009). Reactive components of classroom management refer to the general extent and the adequacy of teacher reactions to acoustic disruptions of the instruction program (e.g., screaming of the students or illicit chatting), motor disruptions (e.g., playing with pens or pencils, eating, chewing gum), or active and passive refusal (e.g., staring out of the window or reading something, which is not relevant for the current classroom activities). More precisely, the term discipline refers either to the degree of disruptions in the classroom or to the teachers' response to these disruptions and their effectiveness to stop misbehavior by the students (Kounin, 2006). These disciplinary responses of a teacher need to be appropriate concerning the type and degree of disruption and correct concerning the addressee (Gold, Förster, & Holodynski, 2013).

Proactive components of classroom management describe the teachers' set of classroom rules and rituals as well as the transparency of consequences for breaking the rules. Rules refer to expectations and standards of behavior in school and, especially, during instruction (Carter & Doyle, 2006; Emmer, Evertson, & Worsham, 2002; Keller, 2010) such as rules in terms of communication and social interactions, student duties, student needs or subject-specific agreements (e.g., rules for safe experimentation in science classes) (Emmer et al., 2002). Oftentimes, rules are regulated in routines or procedures, which refer to rituals (in periodical rhythms) that are provided for starting the lesson on time, giving students the chance to be prepared, or asking for silence in order to regulate the noise level (Carter & Doyle, 2011; Fricke, 2017; Muijs & Reynolds, 2011).

Finally, preventive components of classroom management refer to teacher actions anticipating and preventing disruptions before they compromise the lesson (Borich, 2006). This requires a teacher's general awareness of student activities during the course of the lesson and his or her methods to organize these activities in advance. Key components of preventive classroom management are giving transparency to the students concerning content and methods of instruction, an adequate preparation of learning materials, a constant focus on the complete class, and the realization of a balanced activity level of the students between idleness due to lack of tasks or confusion due to cognitive overload (Borich, 2006; Keller, 2010; Kounin, 2006).

These three components of classroom management skills may be operationalized in three subtests of student ratings termed as ‘discipline’, ‘rules and rituals’, and ‘prevention of disruption’ (Fricke et al., 2012). On a theoretical basis, the latter two components ‘rules and rituals’ and ‘prevention of disruption’ are more closely related, indicating an active, anticipatory, and planned behavior by the teacher (Fricke et al., 2012). However, it is unknown whether this three-dimensional structure of classroom management skills can be supported by empirical evidence when being assessed by means of student ratings.

Aim of this study and research questions

The investigation of the dimensional structure of student responses on classroom management skills is useful to provide educational researchers with differentiated profiles of measures of classroom management skills for studying the effects of these variables in instructional quality models. Thus, the primary aim of the present study was to investigate whether the aforementioned components of classroom management skills can be measured with student ratings with a focus on physics education in a reliable manner. More precisely and referring the three steps of the procedure by Stapleton et al. (2016), three research questions addressed the degree of clustering, the fit of the measurement model, and the estimation of reliability measures:

1. Do student ratings on classroom management skills involve (a) a sufficient level of variability on both levels (ICC(1) > .05; Geldhof et al., 2014) and (b) a sufficient level of agreement among the student ratings from the same classroom (ICC(2) > .50; Stapleton et al. 2016)?
2. Is a three-dimensional profile of classroom management skills with the factors ‘discipline’, ‘rules and rituals’, and ‘prevention of disruption’ (operationalizing reactive, preventive and proactive components of classroom management; Fricke et al., 2012) empirically supported by means of model fit measures against (a) a one-dimensional model and (b) a two-dimensional model where items originally developed for the ‘rule and rituals’ and the ‘prevention of disruption’ subtests are collapsed to one factor, indicating the anticipatory, planning, and structuring nature of the behavioral patterns operationalized by the items from these subtests?
3. Do measures of classroom management skills obtained from the most appropriate measurement model offer a sufficient level of reliability ($\omega > .70$; e.g., Nunnaly, 1978)?

Method

Participants

A data set of *N* = 2680 students (53.28 % females) from 114 classes sampled out of elementary schools (*n* = 1326) and two tracks of secondary schools (lower secondary school, “Hauptschule”, and higher secondary school, “Gymnasium”; *n* = 1354) was analyzed. Their mean age was *M* = 11.18 years (*SD* = 1.15). The 114 teachers (60 % females, 37 % males, 3 % nonresponses) of these classes held a mean teaching experience of *M* = 15.70 years (*SD* = 12.65). The data set was originally collected by the second author in the context of a larger educational research project (Fricke, 2017).

Instrument

A new classroom management questionnaire (“Students Perceptions on Classroom Management”; SPCM) was comprised of seventeen items to assess classroom management skills in physics education. The items were developed according to the framework of reactive (6 items), proactive (5 items), and preventive (6 items) components of classroom management (*c.f.* Fricke, 2017). The verbal stimuli were phrased in a way that they consider physics related instruction in science lessons in German elementary schools and physics lessons in the German secondary school tracks *Hauptschule* and *Gymnasium*. A sample item for the ‘discipline’ subtest is “Students are permanently messing about in our physics lessons.” (item D02; coding inverted), a sample item for the ‘rules and rituals’ subtest is “Classroom rules in physics lessons are well-known to all students.” (item R04a), and a sample item from the ‘prevention of disruption’ subtest is “Our teacher immediately notices students who mess about.” (item P06) (translations from the original German version in Kauertz et al., 2011). The response format was a four-point likert scale ranging from 1 = “I agree” to 4 = “I disagree”.

Statistical approach to data analysis

To investigate the assumed three-dimensional structure of the student ratings on the SPCM questionnaire items by means of a shared cluster construct model, the procedure suggested by Stapleton et al. (2016) with three steps of analyses was applied to the data. With regard to the first research question, ICC(1) and ICC(2) were computed in the first step of analysis for each item of the questionnaire to determine the level of agreement across student ratings from the same classrooms.

With regard to the second research question, an investigation of the dimensional data structure was carried out in the second step of analysis by comparing a unidimensional model (Model 1), a two-dimensional model with items from the ‘discipline’ subtest loading on one factor and items from the ‘rules and rituals’ subtest and the ‘prevention of disruption’ subtest loading on the second factor (Model 2), and a three-dimensional model where items from the 'discipline' subtest, the 'rules and rituals' subtest, and the 'prevention of disruption' subtest each load on one factor (Model 3). Note again that each of the three models includes (unconstrained) covariance estimates between all student ratings on the individual level and that the measurement model was specified only on the cluster level. To determine the most parsimonious measurement model, comparisons of the fit were carried out according to absolute fit statistics and relative fit statistics as described before. Each of the models were estimated by means of a maximum likelihood procedure in a currently unreleased version of lavaan (Rosseel, 2012; see also Huang, 2017, for an alternative estimation procedure in lavaan) incorporating the TL-CFA as part of a multilevel structural equation modeling framework for the R software of statistical computing (R Core Team, 2017).

Finally and with regard to the third research question, composite reliabilities for the correlated factors model as defined in Equation 5 were computed in the third step of analysis based on the between-level estimates by the Excel spreadsheet provided by Cho (2016).

Results

Descriptive statistics of the seventeen items from the SPCM questionnaire are given in Table 1. The student ratings from the rules and rituals scale and the prevention of disruption scale display higher mean scores ranging from 2.81 to 3.36 on the four-point scale compared to the ratings from the discipline scale, which, in fact, have all obtained inverted codings and ranged between 2.49 and 2.66. All items show a substantial level of variation in the student ratings with standard deviations of $SD \geq .90$.

Results concerning the first research question. The investigation of the dimensional structure of student ratings on classroom management skills included the three steps of analyses by Stapleton et al. (2016), beginning with the computation of ICC(1) and ICC(2). The results from the computation of these statistics are also given in Table 1. All items from the questionnaire reached values of the ICC(1) coefficient of at least about .15 and values of the ICC(2) coefficient above .80. Thus, these results demonstrate that a reasonable large and reliable variance proportion is due to shared perceptions of the students, which is a prerequisite for measurement of profiles of teachers' classroom management skills by means of student ratings. The results also satisfy common standards for ICC(1) and ICC(2) for the application of multilevel models (Klein et al., 2000).

Table 1:
Descriptive Statistics Including ICC(1) and ICC(2) of Seventeen Items from the SPCM Questionnaire

Subtest	Item	*M*	*SD*	n_{miss}	ICC(1)	ICC(2)
D	D02	2.64	0.90	161	0.19	0.85
	D04	2.54	0.96	161	0.25	0.89
	D07	2.49	1.04	166	0.20	0.86
	D08	2.58	0.95	165	0.22	0.87
	D09	2.49	1.00	162	0.19	0.85
	D11	2.66	1.05	159	0.20	0.86
R	R02	3.32	0.99	162	0.21	0.86
	R04a	3.24	0.95	163	0.19	0.84
	R10	3.36	0.98	162	0.15	0.80
	R11	2.84	1.31	162	0.45	0.95
	R14	3.28	0.99	158	0.21	0.87
P	P05	3.05	0.97	160	0.15	0.81
	P06	3.23	0.91	160	0.19	0.85
	P09	2.81	0.97	162	0.21	0.86
	P10	2.86	1.06	162	0.15	0.81
	P11a	3.16	0.93	161	0.21	0.86
	P13	2.98	1.03	161	0.21	0.86

Note. n_{mis} = number of missings. D = 'discipline' subtest; R = 'rules and rituals' subtest; P = 'prevention of disruption' subtest.

Results concerning the second research question. Results from the second step of analyses are shown in Table 2. Results from the CFI, the TLI, and, especially, from the $SRMR_B$ index ($SRMR_B$ = 0.257), as a sensitive index for cluster level misspecification, indicate that Model 1, the unidimensional model, does not adequately represent the dimensional structure of the data. Both of Model 2 and Model 3 display acceptable fit according to common cutoff values for absolute fit statistics. In terms of relative fit statistics, the χ^2-test for model comparison and each of the three information criteria AIC, BIC and BIC_{adj} favor Model 3 compared to Model 2. In summary, although no model reaches a preferable $SRMR_B$ fit of less than .08 (Hu & Bentler, 1999), Model 3 incorporating a 'discipline' factor, a 'rules and rituals' factor, and a 'prevention of disruption' factor displays the best fit considering all available statistics. Thus, Model 3 has been selected as measurement model. The cluster level standardized factor loadings and factor correlations in model 3 are given in Figure 2.

Table 2:
Results on Model Comparison Between Three Shared Construct TL-CFA Models for Student Ratings from the SPCM Questionnaire

Model Fit Measure	Model 1	Model 2	Model 3
χ^2-based test of model fit	698.167 (119)	326.624 (118)	281.671 (116)
$p(\chi^2)$	< .001	< .001	< .001
CFI	.946	.981	.985
TLI	.878	.957	.965
RMSEA	.044	.027	.024
SRMR_W	.018	.003	.002
SRMR_B	.257	.102	.095
AIC	101978.040	101608.497	101567.544
BIC	103164.835	102801.110	102771.792
BIC_adj	102516.677	102149.774	102114.101
model comparison with model 1 (χ^2- test)	-	371.54 (1)	416.5 (3)
model comparison with model 1 ($p(\chi^2)$)	-	< .001	< .001
model comparison with model 2 (χ^2- test)	-	-	44.953 (2)
model comparison with model 2 ($p(\chi^2)$)	-	-	< .001

The latent correlations estimated in Model 3 underscored assumptions summarized above of two psychometrically close subdomains and one distinct subdomain of classroom management skills. The two factors of 'rules and rituals' and 'prevention of disruption' were substantially correlated ($r_{\text{R, P}} = .862$), while the 'discipline' factor was only weakly correlated to both the 'rules and rituals' factor ($r_{\text{D, R}} = .\ 171$) and the 'prevention of disruption' factor ($r_{\text{D, P}} = .255$). The strong correlation between the two factors 'rules and rituals' and 'prevention of disruption' also clarifies why the two-dimensional model displayed acceptable model fit in terms of absolute fit statistics.

Results concerning the third research question. In the third step of analyses, between-cluster composite reliabilities were computed for the three factors. Although the number of items was limited, the three factors were found to be very reliable estimates of the three classroom management factors of 'discipline' ($\omega_\text{D} = .970$), 'rules and rituals' ($\omega_\text{R} = .855$), and 'prevention of disruption' ($\omega_\text{P} = .970$).

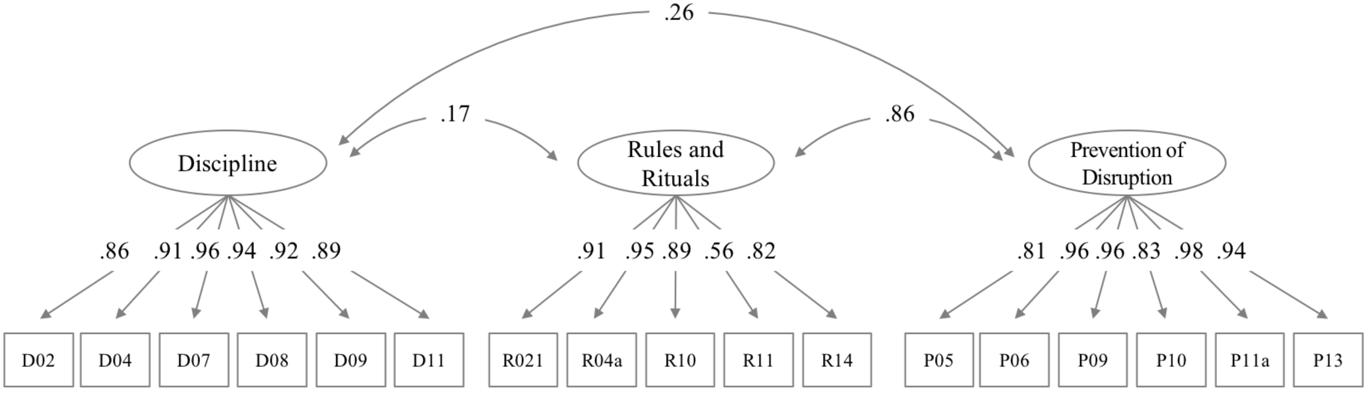

Figure 2:
Three-dimensional cluster level measurement model of classroom management skills (error terms not shown for simplicity)

Discussion

The aim of the present paper was to investigate the dimensional structure of student ratings and differentiate reactive, proactive, and preventive components of classroom management skills to prepare the measurement of profiles of these skills with Physics and Science teachers. To accomplish this aim, the shared cluster construct approach was firstly contrasted in this paper with the configural construct approach and then further characterized in terms of prerequisites of data clustering, model fit, and reliability estimation, according to the procedure by Stapleton et al. (2016). Three components of reactive, proactive and preventive components of classroom management skills were differentiated and, afterwards, three research questions related to the dimensional structure of student ratings on classroom management skills in physics education were derived. Student ratings of classroom management skills of their teachers in physics education were then analyzed in the shared construct TL-CFA to investigate this structure.

Considering the three research questions outlined above, the results from the analyses of student ratings on classroom management skills of teachers in physics education gave some evidence for sufficient agreement on all items among student ratings from the same classroom and for confirming the three-dimensional structure with very reliable measures of reactive, proactive, and preventive components of the construct of interest. As a prerequisite to the application of TL-CFA models for measurement of classroom management skills by student ratings, results concerning the first research question revealed that a large proportion of student ratings on classroom management skills was due to shared perceptions of the construct of interest with ICC(1) and ICC(2) values in line with previous research on instructional quality measures (e.g., Kunter et al., 2008; Wagner et al., 2013).

Results concerning the second research question underscored that reactive, proactive, and preventive behavioral patterns of a teacher can be psychometrically distinguished based on the assessment of student ratings. The proactive and the preventive factor reveal a high degree of overlap, expressed by a substantial (latent) correlation of .862 and best interpreted in terms of an anticipatory, planning, and structuring form of classroom management compared to the rather distinct, reactive form of disciplinary actions. However, from a theoretical point of view, there is consensus that the closely related proactive scale ‘rules and rituals’ and the preventive scale ‘prevention of disruption’ still have to be considered separately (e.g., Borich, 2007). The main difference may originate from the fact that behavioral patterns related to the proactive component not only refer to teacher actions and planning decisions that happen in advance but also refer to a high level of awareness for disruptions, and immediate and adequate reactions to the disruptions by, for example, pointing to the specific classroom rule that was just broken. In contrast, preventive behavior patterns begin to show mainly in advance of a lesson and with the aim of getting the most out of the available time. Minimizing disruptions in advance means “[…] to proceed through lessons smoothly […] with activities and assignments [that] feature stimulating variety and optimal challenge, which help students to sustain their task engagement and minimize disruptions due to boredom or distraction” (Brophy, 2000, p. 11).

The correlations between the reactive factor ‘discipline’ on the one hand, and the proactive factor ‘rules and rituals’ and the preventive factor ‘prevention of disruption’ on the other hand are rather weak, raising the question whether a superordinate concept ‘classroom management’ represents this empirically found correlational structure adequately. Although heterogeneous from an empirical perspective, these three different classroom management skills share the same common goal of maximizing the learning time of their students. Therefore, from the perspective of the authors, the generic term ‘classroom management skills’ in contrast to the more common term of classroom management (which indicates a unitary construct) was used throughout this paper, referring to research on an established construct in instructional quality research but also stressing the heterogeneity of different actions taken by a teacher to maximize the students’ learning time.

Complementing results related to the second research question, findings concerning the third research question also indicated that the three components of classroom management skills were measured in a reliable way in the shared cluster construct model, indicating a psychometrically sound measurement approach to estimate profiles of classroom management skills in school subjects like in Physics or Science. Thus, this study has given some support for a complex dimensional structure of classroom management skills assessed by means of student ratings in Physics or Science compared to the unitary construct definition in most recent studies (e.g., Fauth et al., 2014; Kunter et al., 2008; Wagner et al., 2016).

The assessment of effective classroom management skills is especially important in Physics or Science because the lessons often include experimentation phases, which offer a rather great variety of opportunities to disturb the lesson flow to the students. As these behavioral patterns are directly related to the teachers’ professional vision (Gold et al., 2013), assuring the equivalence of the dimensional structure across classrooms with teachers of different age or job-experience might be object of a next step to investigate the quality of the instrument to measure instructional quality, although this often requires application of different psychometric models. The shared cluster construct model, as a basic TL-CFA model for the statistical analysis of instructional quality (Stapleton et al., 2016), avoids challenges of conceptual nature, like interpreting the construct of interest on the individual level, and of statistical nature, like establishing cross-level invariance. Note that instructional quality measures like classroom management assessed by student ratings typically do not offer a clear interpretation of the construct on the individual level (item variance at individual level might be interpreted as a quantity of unreliability related to the specific method of student self-reports applied to measure classroom management), making the shared cluster construct approach a useful model to investigate reactive, proactive, and preventive components of classroom management in general. However, the investigation of more in-depth research questions related to the equivalence of the measures across different classrooms, teacher groups or even school subjects often requires application of more complex models (like, e.g., the two-level factor model; Wagner et al., 2013). Additionally, further studies concerned with the overlap of the measures to expert ratings based on video recording and to educational outcome variables (e.g., gains in achievement or motivation) also need to underline the usefulness of a profile of classroom management skills in physics education estimated from student ratings. These analyses, however, are beyond the scope of this paper.

References

Borich, G. D. (2006). *Effective teaching methods.* Columbus, Ohio: Pearson Prentice Hall.

Borich, G. D. (2007). *Effective teaching methods. Research-based practice (6th ed.).* Upper Saddle River, NJ: Pearson Education.

Brophy, J. E. (2000). *Teaching* (Educational Practices Series, Vol.1). Brussels: International Academy of Education & International Bureau of Education.

Carter, K. & Doyle, W. (2011). Classroom management in early childhood and elementary classrooms. In C. Evertson & C. Weinstein (Eds.), *Handbook of classroom management: Research, practice, and contemporary issues* (pp. 373–406). Mahwah, NJ: Erlbaum.

Cho, E. (2016). Making reliability reliable: a systematic approach to reliability coefficients. *Organizational Research Methods*, *19*, 651–682. doi:10.1177/1094428116656239

Dollase, R. (2012). *Classroom Management.* Schulmanagement-Handbuch [School management handbook]. München: Oldenbourg.

Doyle, W. (2011). Ecological approaches to classroom management. In C. Evertson & C. Weinstein (Eds.), *Handbook of classroom management: Research, practice, and contemporary issues* (pp. 97–126). Mahwah, NJ: Erlbaum.

Emmer, E. T., Evertson, C. M. & Worsham, M. E. (2002). *Classroom management for secondary teachers* (Vol. 6). Boston, MA: Allyn and Bacon.

Fauth, B., Decristan, J, Rieser, S., Klieme, B., & Büttner, G. (2014). Student ratings of teaching quality in primary school: dimensions and prediction of student outcomes. *Learning & Instruction*, *29*, 1–9. doi:10.1016/j.learninstruc.2013.07.001

Fricke, K. (2017). *Classroom management and its impact on lesson outcomes in physics. A multi-perspective comparison of teaching practices in primary and secondary schools*. Berlin: Logos.

Fricke, K., Ackeren, I. van, Kauertz, A., & Fischer, H. E. (2012). Students‘ perceptions of their teacher’s classroom management in componentary and secondary science lessons and the impact on student achievement. In T. Wubbels, P. den Brok, J. van Tartwijk & J. Levy (Eds.), *Interpersonal relationships in education: An overview of contemporary research* (Advances in learning environments series) (pp. 167-185). Rotterdam, The Netherlands: SENSE Publishers.

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods, 19*, 72–91. doi:10.1037/a0032138

Gold, B., Förster, S., & Holodynski, M. (2013). Evaluation eines videobasierten Trainingsseminars zur Förderung der professionellen Wahrnehmung von Klassenführung im Grundschulunterricht [Evaluating a video-based seminar for university students with the objective of promoting their professional vision of classroom management]. *Zeitschrift für Pädagogische Psychologie, 27*(3), 141–155. doi:10.1024/1010-0652/a000100

Helmke, A. (2009). *Unterrichtsqualität und Lehrerprofessionalität. Diagnose, Evaluation und Verbesserung des Unterrichts* [Quality of instruction and teacher professionalism: Diagnosis, evaluation and improvement of instruction]. Seelze: Knallmeyer.

Hsu, H. Y., Kwok, O. M., Lin, J. H., & Acosta, S. (2015). Detecting misspecified multilevel structural equation models with common fit indices: a monte carlo study. *Multivariate Behavioral Research, 50*, 197–215. doi:10.1080/00273171.2014.977429.

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

Huang, F. L. (2017, September 20). *Conducting multilevel confirmatory factor analysis using R*. Retrieved from: http://faculty.missouri.edu/huangf/data/mcfa/MCFAinRHUANG.pdf

Kauertz, A., Kleickmann, T., Ewerhardy, A., Fricke, K., Lange, K., Ohle, A., … Möller, K. (2011). *Dokumentation der Erhebungsinstrumente im Projekt PLUS* [Technical report of tests and questionnaires in the PLUS project]. Retrieved from http://duepublico.uni-duisburg-essen.de/servlets/DerivateServlet/Derivate-36697/Dokumentation_der_Erhebungsinstrumente_im_Projekt_PLUS_2013_final2.pdf

Keller, G. (2010). *Disziplinmanagement in der Schulklasse: Unterrichtsstörungen vorbeugen – Unterrichtsstörungen bewältigen*, 2. Aufl. [Managing discipline in classes: preventing disruptions—managing disruptions, 2nd ed.]. Bern: Verlag Hans Huber.

Kounin, J. S. (2006). *Techniken der Klassenführung* [Classroom management techniques]. Münster: Waxmann.

Klein, K. J., Bliese, P.D., Kozlowski, S. W. J., Dansereau, F., Gavin, M.B., Griffin, M. A., …, Bligh, M. C. (2000). Multilevel analytical techniques: Commonalities, differences and continuing questions. In K. J. Klein & S.W.J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 512–553). San Francisco, CA: Jossey-Bass.

Klieme, E., Pauli, C., Reusser, K. (2009). The Pythagoras study: Investigating effects of teaching and learning in Swiss and German mathematics classrooms. In T. Janik & T. Seidel (Eds.), *The power of video studies in investigating teaching and learning in the classroom* (pp. 137–160). Münster, Germany: Waxmann.

Kunter, M., Tsai, Y.-M., Klusmann, U., Brunner, M., Krauss, Stefan und Baumert, J. (2008) Students' and mathematics teachers' perceptions of teacher enthusiasm and instruction. *Learning and Instruction*, *18*, 468–482. doi:10.1016/j.learninstruc.2008.06.008

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2x2 taxonomy of multilevel latent contextual model: Accuracy-bias trade-offs in full and partial error correction models. *Psychological Methods*, *16*, 444–467. doi:10.1037/a0024376

Lüdtke, O., Trautwein, U., Kunter, M. & Baumert, J. (2006). Reliability and agreement of student ratings of the classroom environment – A reanalysis of TIMSS data. *Learning Environments Research*, *9*, 215–230. doi:10.1007/s10984-006-9014-8

Marsh, H.W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S., & Köller, O. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist*, *47*, 106–142. doi:10.1080/00461520.2012.670488

Muijs, D. & Reynolds, D. (2011). *Effective teaching: Evidence and practice.* London & Thousand Oaks, CA: Sage.

Nilsen, T. & Gustafsson, J.-E. (ed.) (2016). *Teacher Quality, Instructional Quality and Student Outcomes Relationships Across Countries, Cohorts and Time*. New York: Springer Publishing Company. doi:10.1007/978-3-319-41252-8

Nunnaly, J. (1978). *Psychometric theory*. New York: McGraw-Hill.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing [computer software]. Vienna, Austria. Retrieved from https://www.R-project.org/.

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical Linear Models. Applications and Data Analysis Methods* (2nd Ed.). Thousand Oaks, CA: Sage.

Rosseel, Y. (2012). lavaan: an R package for structural equation modeling. *Journal of Statistical Software*, *48(2)*, 1–36. Retrieved from http://www.jstatsoft.org/v48/i02/

Rosseel, Y. (n.d.). *lavaan: latent variable analysis. Estimators*. Retrieved from http://lavaan.ugent.be/tutorial/est.html

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428. doi:10.1037/0033-2909.86.2.420

Seidel, T. & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*, 454–499. doi:10.3102/0034654307310317

Stapleton L. M., Yang J. S., Hancock G. R. (2016). Construct meaning in multilevel settings. *Journal of Educational and Behavioral Statistics, 41*, 481–520. doi:10.3102/1076998616646200

Wagner, W., Göllner, R., Helmke, A., Trautwein, U., & Lüdtke, O. (2013). Construct validity of student perceptions of instructional quality is high, but not perfect: Dimensionality and generalizability of domain-independent assessments. *Learning and Instruction, 28*, 1–11. doi:10.1016/j.learninstruc.2013.03.003

Wagner, W., Göllner, R., Werth, S., Voss, T., Schmitz, B., & Trautwein, U. (2016). Student and teacher ratings of instructional quality: Consistency of ratings over time, agreement, and predictive power. *Journal of Educational Psychology, 108*, 705–721. doi:10.1037/edu0000075