Peer-reviewed version available at *Soc. Sci.* **2019**, *8*, 46; doi:10.3390/socsci8020046

Research Article

Measurement Invariance of a Direct Behavior Rating Multi Item Scale across Occasions

Markus Gebhardt [1,*], Jeffrey M. DeVries [2], Jana Jungjohann [3], Gino Casale [4], Andreas Gegenfurtner[5] and Tobias Kuhn [6]

[1] Technische Universität Dortmund; Faculty of Rehabilitation Science, Research in Inclusive Education, Dortmund, Germany 1, markus.gebhardt@tu-dortmund.de
[2] Technische Universität Dortmund; Faculty of Rehabilitation Science, Research in Inclusive Education, Dortmund, Germany 2; jeffrey.devries@tu-dortmund.de
[3] Technische Universität Dortmund; Faculty of Rehabilitation Science, Research in Inclusive Education, Dortmund, Germany 3; jana.jungjohann@tu-dortmund.de
[4] Universität zu Köln, Department of Special Education, Germany 4; gino.casale@upb.de
[5] Technische Hochschule Deggendorf, Institut für Qualität und Weiterbildung, Germany 5; andreas.gegenfurtner@th-deg.de
[6] Technische Universität Dortmund; Faculty of Rehabilitation Science, Educational Research Methods, Germany; tobias.kuhn@tu-dortmund.de
* Correspondence: markus.gebhardt@tu-dortmund.de; Tel.: +49-231-755-4546

Abstract: Direct Behavior Rating (DBR) as a behavioral progress monitoring tool can be designed as longitudinal assessment with only short intervals between measurement points. The reliability of these instruments has been evaluated mostly in observational studies with small samples based on generalizability theory. However, for standardized use in the pedagogical field, a larger and broader sample is required in order to assess measurement invariance between different participant groups and over time. Therefore, we constructed a DBR with multiple items to measure the occurrence of specific externalizing and internalizing student classroom behaviors on a Likert scale (1 = never to 7 = always). In a pilot study, two trained raters observed 16 primary school students and rated the student behavior over all items with a satisfactory reliability. In the main study, 108 regular primary school students, 97 regular secondary school students and 14 students in a clinical setting were rated daily over one week (five measurement points). IRT analyses confirmed the instrument's technical adequacy, and latent growth models demonstrated the instrument's stability over time. Further development of the instrument and study designs to implement DBRs are discussed.

Keywords: Direct Behavior Rating 1; Test 2; Sensitivity over time 3; Rating 4; School 5; Classroom Behavior 6; Progress Monitoring 7

1. Introduction

Emotional and behavioral problems in students pose a big challenge in classrooms. These problems have been traditionally structured into externalizing and internalizing behavior problems (Achenbach and Edelbrock 1978). Externalizing behavior problems are outwardly directed behaviors, which are a representation of a maladaptive underregulation of cognitive and emotional states (Achenbach and Edelbrock 1978). Meanwhile, internalizing behavior problems typically develop and persist within an individual, and they represent a maladaptive overregulation of cognitive and emotional states (Achenbach and Edelbrock 1978). According to national and international prevalence studies, 10% to 20% of all school-age children and adolescents show such behavioral problems (e. g., Costello, Mustillo, Erkanli, Keeler and Angold 2003). Longitudinal studies focusing on the consequences of students' externalizing and internalizing behavior problems have shown that the aforementioned behavior patterns in the classroom correlate with academic failure, social exclusion, and delinquency (e.g., Krull, Wilbert and Hennemann 2018, Moffit, Caspi, Harrington and Milne 2002, Reinke, Herman, Petras and Ialongo 2008). In addition, teachers report high levels of stress when they face students' externalizing and internalizing behavior problems in the classroom (e.g., Center and Callaway 1999).

School-based behavioral interventions have been shown to be an efficient way to prevent and decrease the occurrence of externalizing and internalizing behavior problems (e.g., Durlak et al. 2011, Fabiano et al. 2018, Waschbusch et al. 2018). However, the effectiveness of these intervention methods increases when intervention planning, implementation, and evaluation are closely linked to school-based assessment practices (Eklund et al. 2010). Two assessment methods have been shown to lead to more effective interventions (Volpe, Briesch and Chafouleas 2010): universal behavior screening and behavior progress monitoring. Universal screening tools identify students who might benefit from a behavior intervention and additionally guide its planning and implementation. Behavioral progress monitoring is used to evaluate an individual student's response to a behavioral intervention. Behavioral progress monitoring data is collected very frequently up to several times a day. It allows teachers to recognize behavioral changes of the students over a short time period, which assists decisions about maintaining or modifying the intervention.

Although many existing tools can be used for universal behavior screening (e.g., Daniels, Volpe, Briesch and Fabiano 2014, Volpe et al. 2018 for an overview), the development of methods that can be used for behavioral progress monitoring is still in its initial stages. Traditionally, there are two widely used approaches that have been used for school-based behavior assessment: behavior rating scales (BRS) and systematic direct behavior observations (SDO; Christ, Riley-Tillman and Chafouleas 2009). BRS usually consist a pool of items representing specific behaviors that an individual might exhibit. The intensity or frequency of these behaviors are rated on a Likert scale. Therefore, the documentation and interpretation of the behavior occur at the same time. BRS can be completed by multi-informants such as the teachers, the parents, or the individual itself. BRS are an efficient way to measure specific behaviors, since they are easy to understand, complete, and interpret. However, the scores generated by BRS represent a subjective perception of an individual's behavior. SDO, in contrast, represent an objective tool to assess a student's behavior (Volpe, DiPerna, Hintze and Shapiro 2005). In SDO, the documentation and interpretation of an individual's behavior are usually separated. The observer identifies and defines the behavior of interest, and the observation interval. Afterwards, the targeted behavior will be observed in the relevant interval by using time-sampling methods. Finally, the observation scores are analyzed and interpreted. While this procedure generates objective, reliable, and valid data, it is work and time intensive. Furthermore, observation training is often required. In conclusion, BRS and SDO alone have limitations when collecting behavioral progress monitoring data (Christ et al. 2009).

Direct Behavior Rating (DBR) represents a relatively new assessment method, which allows for progress monitoring measurements over short intervals. DBR is a hybrid form of systematic direct observation and behavior rating scales wherein individuals observe and rate (e.g., on a Likert scale) a behavior in a specific situation immediately afterwards (Chafouleas 2011). In recent years, two DBR forms have been developed and evaluated for progress-monitoring purposes: Single-Item Scales (DBR-SIS) and Multi-Item Scales (DBR-MIS; Volpe and Briesch 2015). DBR-SIS usually targets more global behaviors (e.g. academically engagement, disruptive behavior) and may be the most efficient way to broadly measure a student's overall level of behavioral success. This information could be useful when a student exhibits a broad range of specific problem behaviors that are related to problem behavior in general. However, DBR-SIS have not typically been used to assess specific classroom behaviors (e.g., hand raising), which might be more informative for evaluating a student's response to behavioral intervention. In contrast, DBR-MIS usually include three to five specific behavior items (e.g. completes classwork in allowed time, starts working independently, turns in assignments appropriately) that operationalize a higher-order behavioral dimension. These more specific items can then be analyzed individually or added up to produce a sum score (Volpe and Briesch 2012).

Previous studies have shown that DBR meets the criteria required for behavioral progress-monitoring. First, DBR is feasible and effective because it does not require extensive materials and the ratings can be completed easily in a few minutes (Chafouleas 2011). Second, DBR is flexible because a broad range of observable behaviors (at both the global and specific levels) can be addressed. Third, DBR is repeatable because the same behavior target can be observed and rated across many observations. Fourth, the psychometric quality of DBR has been supported by a broad range of evaluation studies focusing on the performance of the tool under different measurement conditions (Chafouleas 2011; Christ et al. 2009; Huber & Rietz, 2015).

Most DBR studies have assessed reliability using Generalizability Theory (GT; see Huber & Rietz, 2015). Within generalizability theory, which represents a liberalization of Classical Test Theory (CTT), assessments are tied closely to the target populations with respect to the variability of the targeted behaviors. This technique can establish the external validity of a DBR by ensuring that the behavioral targets and evaluation groups are well matched. Most studies were designed along generalizability theory in order to measure the true behavior and investigate potential factors (and their interactions) that might influence the variance in the generated scores (e.g., such as multiple raters and multiple time points). Such studies are necessary to examine the reliability of behavioral assessment and to determine conditions that might increase the reliability (Cronbach, Gleser, Nanda and Rajaratnam 1972). Previous studies found that DBR generates reliable scores by reflecting a large amount of variance explained by the actual student's behavior (e.g., Owens and Evans 2017). However, results from different raters across multiple time points indicate that different persons rate the same behavior differently and that students behave differently across multiple occasions (e.g., Briesch, Chafouleas and Riley-Tillman 2010; Volpe and Briesch 2012, Briesch, Volpe and Ferguson 2014). Therefore, multiple measurement points are necessary to provide a stable score that is still interpretable. Previous generalizability studies showed that valid results are generated within 4 to 20 measurement points, and that fewer measurement points were needed when DBR-MIS were used (e.g., Casale, Hennemann, Volpe, Briesch and Grosche 2015; Volpe, Briesch and Gadow 2011, Volpe and Briesch 2012).

Even if the results on the psychometric characteristics of DBR are promising, there are still two remaining issues. First, most of the previous studies had small sample groups with five to ten students and three or more raters. Because of theoretical assumptions in generalizability studies, smaller samples are often used, but such a small sample is insufficient for the evaluation of internal validity and testing technical adequacy of the test itself. For instance, Rasch modelling may require 100 participants, with 250 for high stakes decisions like screenings, diagnoses, or classroom advancement, to obtain sufficiently precise parameter estimates (Linacre 1994). It is therefore important to embed evaluation throughout the development process and to use an evaluation sample that is sufficient size for IRT analyses. Therefore, DBR should be developed in lines of both generalizability theory and IRT, with evaluation embedded throughout development.

Second, measurement invariance of DBR across multiple occasions has yet not been examined. Since DBR was developed for assessment within a problem-solving model, it must be sensible of behavioral progress (Deno 2003, Good and Jefferson 1998). Only when DBR scores are comparable over time, can the results be used to draw valid conclusions regarding the behavioral progress and responses to behavioral interventions (Gebhardt, Heine, Förster and Zeuch 2015).

2. Present Study

We designed a DBR based on the principles of an established screening instrument for use in the educational field. While single teacher ratings can be biased, they are still standard practice in educational evaluations. In order to ensure the psychometric quality of our DBR, we used five

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 October 2018 doi:10.20944/preprints201810.0455.v1

Peer-reviewed version available at Soc. Sci. 2019, 8, 46; doi:10.3390/socsci8020046

measurement points per single rater. Teacher ratings were used to allow for this relatively high number of ratings required per student.

We analyzed our new instrument with an IRT Rasch model and then evaluated the measurement invariance based on gender, migration background, and school level separately. We also used a latent growth model to investigate systematic changes in ratings over time based on these qualities.

3. Materials and Methods

The analyses were carried out with the statistics program R (R Core Team 2013) using the package pairwise (Heine 2014). Here the method of explicit calculation of the item parameters in the fast model was used by the pairwise item comparison (Choppin 1968, Wright and Masters 1982). This method is particularly suitable for determining the sample-invariant item parameters for the calibration of a given item pool for a unidimensional model (Choppin 1968). The pairwise estimator is also suitable for small samples or data sets with missing values (Wright and Masters 1982, Heine and Tarnai 2015, Heine, Gebhardt, Schwab, Neumann, Gorges and Wild 2018). First, the measurement invariance over the four time points was checked by means of the graphical model test. Second, the item parameters were calculated over all five measurement times and third, the fit of the model was determined at all measurement times using mean square fit statistics (infit and outfit). The personal parameters were estimated for the respective measurement times using the weighted maximum likelihood method (WLE, Warm 1989). For the common item parameters, the point-biserial correlations with the scale value (WLE estimator) are reported as selectivity for the respective measurement time.

3.1. Sample

219 students were rated by 41 teachers. Teachers had the instruction to choose five students with external or internal behavior problems in their classes. On average, the teachers rated five students. Therefore, our sample is not a full class sample, but consists only of students with behavioral problems. 108 primary school students, 97 secondary students and 14 students in a clinical setting were rated over one week (five measurement points). For each student, the same teacher provided the rating at each measurement point. In the school sample 35 students (17%) had been diagnosed with special education needs, 40 students had a migration background (21%) and 145 students were boys (83%). In the clinical sample 5 students had a migration background (36%) and 11 students were boys (79%).

3.2. Instrument

We developed multidimensional DBR-MIS with different dimensions for externalizing, internalizing, and positive behaviors (Gebhardt, de Vries, Jungjohann & Casale, 2018). Along the lines of other established screening tools (i.e., the Strengths and Difficulties Questionnaire; Goodman 1997, 2001, Voss and Gebhardt 2017, DeVries, Voss, and Gebhardt, 2018), we created a six-dimensional scale divided into 3 areas, which were internalizing problems, externalizing problems, and positive behaviors in school. Internalizing problems included the dimensions depressive and anxious behaviors (DAB) and social interaction problems (SIP). Externalizing problems included the dimensions disruptive behavior (DB) and academic engagement problems (AEP). Lastly, positive behaviors in school included the dimensions scholastic behavior (SB) and prosocial behavior (PS).

We constructed three new items per dimension for this DBR. We adapted the items to measure a single behavior, which can be observed in one school hour. We constructed the dimensions disruptive behavior, academic engagement problems, depressive and anxious behaviors, social interactions problems referring to the Strength and Difficulties Questionnaire (SDQ; Goodman et al. 2010) and used the wording from the ICD 10. All items had seven categories from 1 (never) to 7 (always), which was suggested by Christ, Riley-Tillman and Chafouleas (2009). Additionally, we constructed a new scale about school behavior. All items were unidirectional.

Peer-reviewed version available at *Soc. Sci.* **2019**, *8*, 46; doi:10.3390/socsci8020046

In an initial exploratory study, we asked three special school teachers in special schools which were attended by children with emotional and social problems to provide an expert-review of the items. Each special school teacher rated three children with emotional and social problems. Afterwards, we interviewed the teachers and adapted the items. Next, we tested the new items in a second pilot study with 15 students and two trained raters. The raters had a high compliance rate of spearman's rho = .84. Finally, the items were discussed with raters and minor rewording was done. All items are presented in the appendix in both English and German language. The questionnaire is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike International License 4.0 (Gebhardt, de Vries, Jungjohann & Casale, 2018).

4. Results

4.1. Measurement invariance between time points one and five

First, graphical model tests confirmed the measurement invariance between time points one and five separately for every dimension (see Figure 1). For this analysis, response categories six and seven were combined, because the category seven was rarely used.

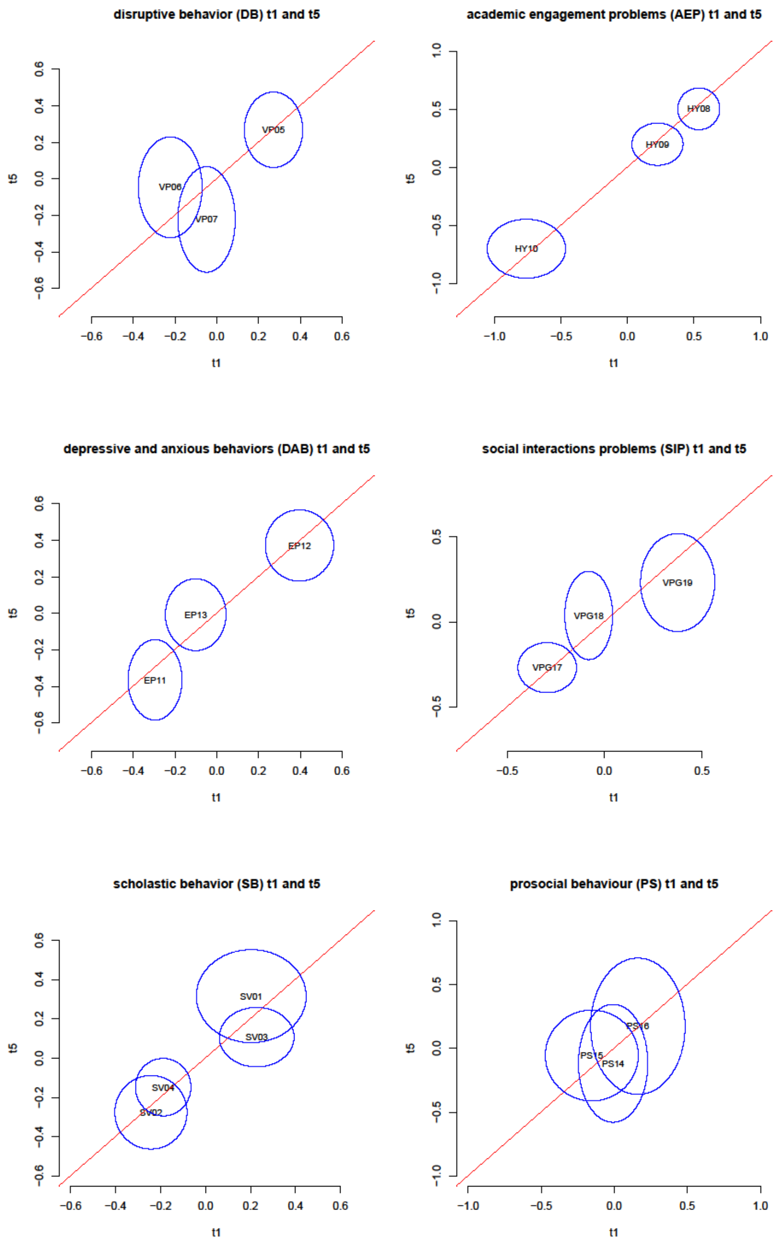

Figure A1. Graphical Model-Test with split criterion time points one and five for six scales disruptive behavior (DB), academic engagement problems (AEP), depressive and anxious behaviors (DAB), social interactions problems (SIP), scholastic behavior (SB) and prosocial behavior (PS) of the DBR

In addition, we investigated the model at the item level via infit and outfit. The outfit is the sum of squared standardized residuals and is sensitive to raters and oversights. In contrast, the infit is weighted for information and shows distortions in the sample (such as Guttman Pattern). The results from the root-mean-square statistics (INFIT and OUTFIT) as well as the point-biserial correlation coefficients are presented in Table 1. They indicate that most of the items fit well with the assumptions of the Rasch model (Wright and Masters 1982). Only the item SB13 "Participates in class" shows a low discrimination. Nonetheless, it is still in the range between 0.5 and 1.5, which Linacre (2002) has suggested as an acceptable range for questionnaires. Although for high-stake tests used to evaluate students, a stricter range of 0.8 to 1.2 is proposed (Wright and Linacre 1994). The WLE reliability,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 October 2018 doi:10.20944/preprints201810.0455.v1

Peer-reviewed version available at Soc. Sci. 2019, 8, 46; doi:10.3390/socsci8020046

which is comparable to Cronbachs Alpha, was sufficient (DB 0.76, AEP 0.77, DAB 0.70, SIP 0.65, SB 0.77, PS 0.87).

Table A1. Stability of the categories in a model over all five times are shown with the point-biserial correlation coefficients and the root-mean-square statistics (INFIT and OUTFIT)

Items	Response level						rout-mean-square statistics	
	1	2	3	4	5	6	$Infit_{MSQ}$	$Outfit_{MSQ}$
DB01	-0.70	-0.11	0.13	0.23	0.40	0.62	0.97	1.00
DB02	-0.62	-0.22	0.04	0.26	0.33	0.62	1.07	1.06
DB03	-0.70	-0.14	0.04	0.21	0.37	0.66	0.96	0.94
AEP04	-0.67	-0.09	-0.01	0.16	0.29	0.58	1.05	1.09
AEP05	-0.62	-0.28	-0.06	0.16	0.36	0.53	1.05	1.05
AEP06	-0.54	-0.39	-0.25	-0.03	0.20	0.68	0.90	0.85*
DAB07	-0.70	-0.18	0.02	0.25	0.23	0.54	0.97	0.92
DAB08	-0.67	0.10	0.08	0.21	0.29	0.42	1.07	1.09*
DAB09	-0.70	0.04	0.08	0.36	0.28	0.45	0.97	0.97
SIP10	-0.65	-0.16	0.01	0.25	0.35	0.63	1.03	1.00
SIP11	-0.71	0.00	0.18	0.29	0.41	0.57	0.96	0,96
SIP12	-0.67	-0.10	0.06	0.29	0.38	0.59	1.01	1,05
SB13	-0.41	-0.14	0.01	0.06	0.35	0.39	1.49*	1,49*
SB14	-0.49	-0.30	-0.17	0.08	0.26	0.55	0.97	0,98
SB15	-0.55	-0.36	-0.04	0.20	0.46	0.51	0.74	0,72
SB16	-0.48	-0.35	-0.17	0.10	0.34	0.59	0.80	0,80
PS17	-0.62	-0.34	-0.16	0.10	0.40	0.56	1.02	1.04
PS18	-0.63	-0.36	-0.16	0.10	0.37	0.61	0.94	0.92
PS19	-0.63	-0.33	-0.13	0.17	0.38	0.58	1.04	1.04

Note: *$p < 0.05$

4.2. Measurement invariance at measurement point five

A CFA tested the hypothesized 6-factor structure using data from measurement point five. Initial results produced an unacceptable initial fit with RMSEA = 0.10, CFI = 0.87, and SRMR = 0.09. However, a modified model produced an acceptable fit, with RMSEA =0.08, CFI = 0.92, and SRMR = 0.06. These minor modifications included dropping item SB13 because of a low loading (0.34), and allowing item SP11 to cross load onto DB.

Measurement invariance was also assessed at measurement point 5 across gender, migration background, and school level. For the school level assessment only, the clinical population was excluded. All analyses of invariance were conducted in Mplus 7.4. We assessed weak invariance by

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 October 2018 doi:10.20944/preprints201810.0455.v1

Peer-reviewed version available at Soc. Sci. 2019, 8, 46; doi:10.3390/socsci8020046

comparing the fits of the configural to metric models and strong invariance by comparing fits of the metric and scalar models (see the procedure described by Dimitrov 2017). We used the threshold of ΔCFI < .010 to indicate a significant change in fit.

Results upheld both weak invariance in all cases with ΔCFI < .010. Similarly, strong invariance was found for gender and migration background, ΔCFI < .010. However, strong invariance was not upheld for school level, ΔCFI = .012. We proceeded to test for partial invariance by individual freeing intercepts with the greatest effect on $\chi 2$ until the difference between the metric and scalar model was under threshold of ΔCFI < .010. We used the standard for partial invariance of fewer than 20% of freed intercepts and loadings (see Levine et al. 2003). The freeing of a single intercept (AEP 05) resulted in a net ΔCFI = .005, under the threshold of .010. As this represented only 3% of the model's intercepts and loadings, we concluded that the instrument possessed sufficient partial invariance across grade level, and that overall comparisons across gender, migration background, and grade level are meaningful with the instrument.

4.3. Latent Growth Models

Four separate latent growth models were calculated to estimate the change in individual WLE scores over time. In these models disruptive behavior (DB) and academic engagement problems (AEP) were collapsed into a single externalizing factor, and depressive and anxious behaviors (DAB) and social interactions problems (SIP) were collapsed into a single internalizing factor. In the models, gender, migration background, attending the secondary school, and attending the clinical school were used as predictors for the slope and intercept parameters in these models. As Table 2 describes, model fits were good in all cases. Table 3 describes the intercept and slope values for the models, as well as the path loadings. Overall slopes were insignificant, but overall intercepts differed from zero. The distribution of slopes can be seen in Figure 2, where the slopes are all close to zero. This indicates that individual persons did not change much on average over time. Furthermore, boys had higher levels of externalizing than girls, and children with a migration background also had a higher level of externalizing behavior. The clinical subsample showed a higher slope of internalizing as well as a lower slope in scholastic behavior, indicating more problematic development within a short time frame. However, the clinical subsample displayed a higher intercept in scholastic behavior and internalizing. Girls had higher scores on prosocial behavior and scholastic behavior. The clinical population displayed a significant effect on the slopes for internalizing, scholastic behavior and prosocial behavior. This may indicate a systematic change in these values for this population.

Table A2. Fit values for the Latent Growth Models

Dimension	RMSEA (90% CI)	CFI	SRMR
Externalizing (AEP+DB)	0.00 (0.00 - 0.01)	1.00	0.02
Internalizing (DAB+SIP)	0.00 (0.00 - 0.03)	1.00	0.01
Prosocial Behavior	0.06 (0.02 - 0.09)	0.99	0.03
Scholastic Behavior	0.00 (0.00 - 0.04)	1.00	0.02

Table A3. Relevant values from latent growth models

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 October 2018 doi:10.20944/preprints201810.0455.v1

Peer-reviewed version available at Soc. Sci. 2019, 8, 46; doi:10.3390/socsci8020046

Dimension	Model	Intercept	Slope
Externalizing (AEP + DB)	Overall	-2.02***	-0.03
	Gender	0.95*	-0.02
	Migration Status	0.48*	0.03
	Secondary School	0.08	-0.01
	Clinical School	0.18	0.06
Internalizing (DAB+SIP)	Overall	-0.61*	-0.03
	Gender	-0.01	-0.04
	Migration Status	0.00	0.05
	Secondary School	0.24	0.06*
	Clinical School	0.43***	0.08****
Prosocial Behavior	Overall	1.09***	0.02
	Gender	-1.42***	0.08
	Migration Status	-0.55	-0.06
	Secondary School	-0.2	-0.07
	Clinical School	-0.03	-0.13*
Scholastic Behavior	Overall	2.63***	-0.03
	Gender	-0.74***	0.01
	Migration Status	-0.35	0.01
	Secondary School	-0.04	-0.01
	Clinical School	0.51*	-0.13**

Note: *significant at $p < .05$

** significant at $p < .01$

***significant at $p < .001$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 October 2018 doi:10.20944/preprints201810.0455.v1

Peer-reviewed version available at Soc. Sci. 2019, 8, 46; doi:10.3390/socsci8020046

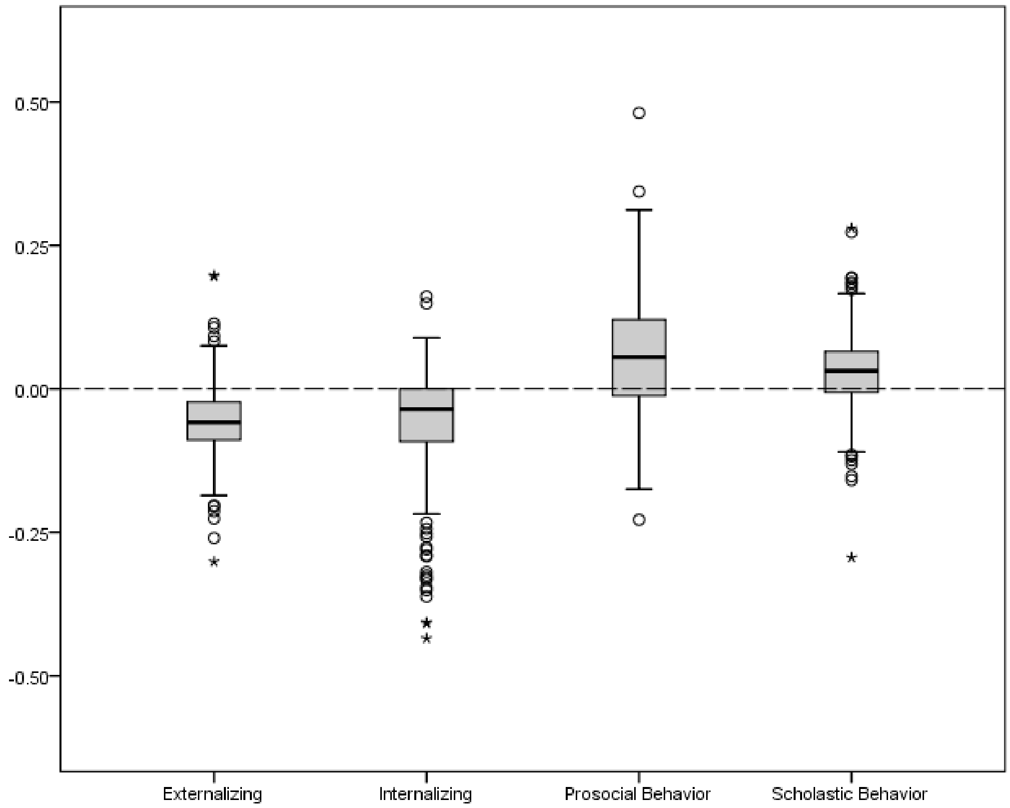

Figure 2. Individual unstandardized slopes for the latent growth models

4.4. Intraclass Correlations of diferent raters

The intraclass correlations based on rater remained low for the same 4 subscales at measurement point five, ICCexternalizing < .01, ICCinternalizing = .07, ICCprosocial behavior = .01, and ICCscholastic behavior < .01.

5. Discussion

This study demonstrated an approach that complements generalizability theory in order to examine the item characteristics and the stability of DBR ratings across time points. Our relatively large sample allowed more detailed IRT analysis than the smaller samples of generalizability theory-based DBR development and assessment. This represents a significant addition to previous DBR assessment and development techniques.

Our DBR-MIS had a high compliance rate by two trained raters in a pilot study and showed in IRT analyses invariance over five measurement points and satisfactory reliability on the item level. Results from the CFA confirmed the overall factor structure parallels the structure of past research (e.g., SDQ; Goodman et al. 2010), with only minor modifications. Invariance tests revealed its applicability to diverse groups based on gender, migration background, and school level. Results of the latent growth models confirmed the overall stability of the scores across all five measurement points. The intraclass correlations based on rater indicate that there was little effect of rater bias on the overall WLE scores.

We did not provide any experimental treatments, and every teacher rated their own students individually. Therefore, it was expected that there would be no difference between school levels or types or across measurement points. Most teachers provided ratings that were in the center category and their ratings remained stable. The highest category, seven (always), was so rarely used, that we needed to combine this category with a lower rating for the IRT model. Therefore, our instrument cannot compare different groups of students to one measurement point, but it is instead more suited

Peer-reviewed version available at Soc. Sci. 2019, 8, 46; doi:10.3390/socsci8020046

to measure the individual change over time. Additional studies are needed to measure the sensitivity to change in behavior over time. For this, studies with interventions are needed to explain the behavior over short and long time periods. In addition, our design did not allow for a detailed analysis of rater effects (e.g., rater severity, rater drift over time), which can affect longitudinal ratings substantially. To disentangle rater effects from item stability/sensitivity to interventions over time, more complex deigns are especially needed.

6. Conclusion

It is possible to develop direct behavior ratings methods in a rigorous manner. This technique provides the advantages of IRT analyses, such as systematic reliability analyses, to the area of behavior rating. We investigated behavioral measurements containing constant items over short measurement periods, which were also by the same teacher at each measurement point. These measurements are stable and individually evaluable. More work to measure the sensitivity to change and responses to treatment is needed to further develop these methods.

Author Contributions: conceptualization, M.G. and G.C.; methodology, M.G.; validation, M.G., J.M.D. and J.T.K.; formal analysis, M.G. and J.M.D.; investigation, M.G. and J.J.; resources, M.G., J.M.D and J.J.; data curation, M.G. and J.M.D.; writing—original draft preparation, M.G., J.M.D. and G.C.; writing—review and editing, M.G., J.J., A.G., J.T.K.; visualization, J.J.; supervision, M.G. and J.T.K.; project administration, M.G.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Ethical Statement: All subjects gave their informed consent for inclusion before they participated in the study. The study was conducted in accordance with the dean of the Faculty of Rehabilitation Science, Technical University of Dortmund. An additional ethics approval was not required for this study as per Institution's guidelines and national regulations..

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 October 2018 doi:10.20944/preprints201810.0455.v1

Peer-reviewed version available at *Soc. Sci.* **2019**, *8*, 46; doi:10.3390/socsci8020046

Appendix A

Instructions and procedures for the "Questionnaire for Monitoring Behavior in Schools" (QMBS)

This questionnaire covers the behavior of pupils during a clearly defined timeframe (e.g., one day of instruction). The goal is to get subsequent ratings in comparable time frames (e.g., a math lesson or the whole day). With such a series, it is possible to map the pattern of behavior of the pupil.

You are free to decide upon the observation timeframe, except that they are comparable in terms of length and didactic content. Also, the same person should rate the same pupil at each timeframe.

If you encounter items that you can not rate based on observations within the timeframe, check the value of the previous day or leave this item blank.

Peer-reviewed version available at Soc. Sci. 2019, 8, 46; doi:10.3390/socsci8020046

Questionnaire for Monitoring Behavior in Schools (QMBS)

Nr.	Items	Never						Always
	Externalizing Behavior							
	Disruptive Behavior (DB)							
1	Has temper tantrums or a hot temper, has a low frustration tolerance.	1	2	3	4	5	6	7
2	Disobeys rules and does not listen to the teacher	1	2	3	4	5	6	7
3	Argues with classmates/provokes classmates with his/her behavior	1	2	3	4	5	6	7
	Academic Engagement Problems (AEP)							
4	Fidgets or squirms, is restless/overactive	1	2	3	4	5	6	7
5	Frequently quits tasks early	1	2	3	4	5	6	7
6	Easily distracted	1	2	3	4	5	6	7
	Internalizing Behavior							
	Depressive and anxious behaviors (DAB)							
7	Seems worried, sad, or depressed	1	2	3	4	5	6	7
8	Seems Fearful	1	2	3	4	5	6	7
9	Seems nervous.	1	2	3	4	5	6	7
	Social interactions problems (SIP)							
10	Works/plays mostly alone, prefers to be alone	1	2	3	4	5	6	7
11	Teased or bullied by classmates, easily provoked	1	2	3	4	5	6	7
12	Gets along better with adults than with other children	1	2	3	4	5	6	7
	Positive Behavior in School							
	Scholastic Behavior (SB)							
13	Participates in class	1	2	3	4	5	6	7
14	Follows rules for speaking in class (i.e., raises hand)	1	2	3	4	5	6	7
15	Concentrates on his/her schoolwork	1	2	3	4	5	6	7
16	Works quietly at his/her desk/does not refuse assignments	1	2	3	4	5	6	7
	Prosocial Behavior (PS)							
17	Considerate of other people's feelings	1	2	3	4	5	6	7
18	Helpful to others	1	2	3	4	5	6	7
19	Cooperative in partner and group situations	1	2	3	4	5	6	7

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 October 2018 doi:10.20944/preprints201810.0455.v1
Peer-reviewed version available at Soc. Sci. 2019, 8, 46; doi:10.3390/socsci8020046

Instruktionen zur Durchführung des „Fragebogens zur Verhaltensdiagnostik in der Schule“

Dieser Fragebogen bzw. das Direct Behavior Rating (DBR) bezieht sich auf das Verhalten der Schülerin bzw. des Schülers im Klassenraum während eines klar umgrenzten Zeitraums (z. B. eines Unterrichtstags). Ziel ist es, mehrere aufeinanderfolgende Verhaltenseinschätzungen für einen vergleichbaren Zeitraum (z. B. die Mathematikstunde oder den gesamten Unterrichtstag) zu bekommen. Eine solche Zeitreihe an Daten kann dann den Verlauf des Verhaltens einer Schülerin/ eines Schülers abbilden.

Sie können die Beobachtungszeiträume frei wählen, sollten jedoch darauf achten, dass sie vergleichbar hinsichtlich Länge und methodisch-didaktischer Aufbereitung sind. Außerdem sollte immer die gleiche Person den DBR ausfüllen.

Sollten Sie auf Items stoßen, die Sie auf Basis des von Ihnen beobachteten Zeitraums nicht bewerten können, kreuzen Sie bitte den Wert des Vortages an/lassen Sie dieses Item bitte aus.

doi:10.20944/preprints201810.0455.v1

Peer-reviewed version available at Soc. Sci. 2019, 8, 46; doi:10.3390/socsci8020046

Fragebogen zur Verhaltensdiagnostik in der Schule

Nr.	Items	Nie						Immer
	Externalisierendes Verhalten							
	Störendes und auflehnendes Verhalten (SAV)							
1	Verhält sich wütend und aufbrausend	1	2	3	4	5	6	7
2	Missachtet Regeln und hört nicht auf die Lehrkraft	1	2	3	4	5	6	7
3	Streitet sich mit Mitschüler_innen/provoziert durch eigenes Verhalten seine Mitschüler_innen	1	2	3	4	5	6	7
	Verhaltensprobleme beim Lernen (VPL)							
4	Zappelt, ist (motorisch) unruhig/ überaktiv	1	2	3	4	5	6	7
5	Bricht Aufgaben häufig früh ab	1	2	3	4	5	6	7
6	Lässt sich schnell und leicht ablenken	1	2	3	4	5	6	7
	Internalisierendes Verhalten							
	Depressives und ängstliches Verhalten (DAV)							
7	Wirkt besorgt, betrübt oder bedrückt	1	2	3	4	5	6	7
8	Wirkt ängstlich/ fürchtet sich	1	2	3	4	5	6	7
9	Wirkt nervös (sucht Nähe zu Erwachsenen)	1	2	3	4	5	6	7
	Probleme in sozialen Interaktionen (PSI)							
10	Arbeitet/spielt meist alleine	1	2	3	4	5	6	7
11	Wird von Mitschüler_innen gehänselt oder geärgert, lässt sich provozieren	1	2	3	4	5	6	7
12	Arbeitet/spielt häufiger mit Erwachsenen als mit Mitschüler_innen	1	2	3	4	5	6	7
	Positives Schulverhalten							
	Schulbezogenes Verhalten (SV)							
13	Meldet sich im Unterricht	1	2	3	4	5	6	7
14	hält sich an Gesprächsregeln	1	2	3	4	5	6	7
15	Richtet Aufmerksamkeit/Konzentration auf die Bearbeitung der Aufgabe	1	2	3	4	5	6	7
16	Arbeitet ruhig am Platz und verweigert nicht die Mitarbeit	1	2	3	4	5	6	7
	Prosoziales Verhalten (PS)							
17	Verhält sich anderen gegenüber rücksichtsvoll	1	2	3	4	5	6	7
18	Verhält sich anderen gegenüber hilfsbereit	1	2	3	4	5	6	7
19	Verhält sich in Partner- und Gruppensituationen kooperativ	1	2	3	4	5	6	7

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 October 2018 doi:10.20944/preprints201810.0455.v1

Peer-reviewed version available at *Soc. Sci.* **2019**, *8*, 46; doi:10.3390/socsci8020046

Appendix B

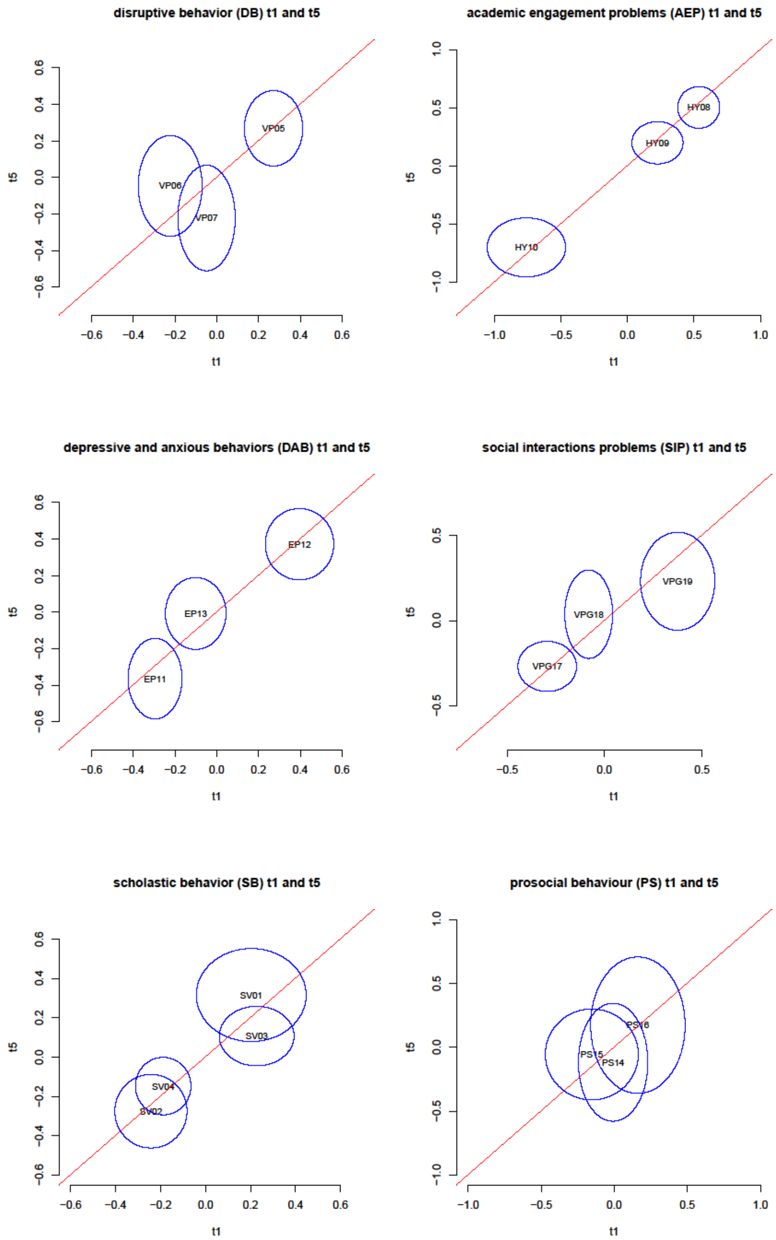

Figure A1. Graphical Model-Test with split criterion timepoints 1 and 5 for six scales disruptive behavior (DB), academic engagement problems (AEP), depressive and anxious behaviors (DAB), social interactions problems (SIP), scholastic behavior (SB) and prosocial behavior (PS) of the DBR

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 October 2018 doi:10.20944/preprints201810.0455.v1

Peer-reviewed version available at *Soc. Sci.* **2019**, *8*, 46; doi:10.3390/socsci8020046

Table A1. Stability of the categories in a model over all five times are shown with the point-biserial correlation coefficients and the rout-mean-square statistics (INFIT and OUTFIT)

Items	Response level						rout-mean-square statistics	
	1	2	3	4	5	6	$Infit_{MSQ}$	$Outfit_{MSQ}$
DB01	-0.70	-0.11	0.13	0.23	0.40	0.62	0.97	1.00
DB02	-0.62	-0.22	0.04	0.26	0.33	0.62	1.07	1.06
DB03	-0.70	-0.14	0.04	0.21	0.37	0.66	0.96	0.94
AEP04	-0.67	-0.09	-0.01	0.16	0.29	0.58	1.05	1.09
AEP05	-0.62	-0.28	-0.06	0.16	0.36	0.53	1.05	1.05
AEP06	-0.54	-0.39	-0.25	-0.03	0.20	0.68	0.90	0.85*
DAB07	-0.70	-0.18	0.02	0.25	0.23	0.54	0.97	0.92
DAB08	-0.67	0.10	0.08	0.21	0.29	0.42	1.07	1.09*
DAB09	-0.70	0.04	0.08	0.36	0.28	0.45	0.97	0.97
SIP10	-0.65	-0.16	0.01	0.25	0.35	0.63	1.03	1.00
SIP11	-0.71	0.00	0.18	0.29	0.41	0.57	0.96	0,96
SIP12	-0.67	-0.10	0.06	0.29	0.38	0.59	1.01	1,05
SB13	-0.41	-0.14	0.01	0.06	0.35	0.39	1.49*	1,49*
SB14	-0.49	-0.30	-0.17	0.08	0.26	0.55	0.97	0,98
SB15	-0.55	-0.36	-0.04	0.20	0.46	0.51	0.74	0,72
SB16	-0.48	-0.35	-0.17	0.10	0.34	0.59	0.80	0,80
PS17	-0.62	-0.34	-0.16	0.10	0.40	0.56	1.02	1.04
PS18	-0.63	-0.36	-0.16	0.10	0.37	0.61	0.94	0.92
PS19	-0.63	-0.33	-0.13	0.17	0.38	0.58	1.04	1.04

Note: $*p < 0.05$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 October 2018 doi:10.20944/preprints201810.0455.v1

Peer-reviewed version available at *Soc. Sci.* **2019**, *8*, 46; doi:10.3390/socsci8020046

Table A2. Fit values for the Latent Growth Models

	RMSEA (90% CI)	**CFI**	**SRMR**
Externalizing (HY+CP)	0.00 (0.00 - 0.01)	1.00	0.02
Internalizing (EP+PP)	0.00 (0.00 - 0.03)	1.00	0.01
Prosocial Behavior	0.06 (0.02 - 0.09)	0.99	0.03
Scholastic Behavior	0.00 (0.00 - 0.04)	1.00	0.02

Table A3. Relevant values from latent growth models

		Intercept	**Slope**
Externalizing (HY+CP)	Overall	-2.02***	-0.03
	Gender	0.95*	-0.02
	Migration Status	0.48*	0.03
	Secondary School	0.08	-0.01
	Clinical School	0.18	0.06
Internalizing (EP+PP)	Overall	-0.61*	-0.03
	Gender	-0.01	-0.04
	Migration Status	0.00	0.05
	Secondary School	0.24	0.06*
	Clinical School	0.43***	0.08****
Prosocial Behavior	Overall	1.09***	0.02
	Gender	-1.42***	0.08
	Migration Status	-0.55	-0.06
	Secondary School	-0.2	-0.07
	Clinical School	-0.03	-0.13*
Scholastic Behavior	Overall	2.63***	-0.03
	Gender	-0.74***	0.01
	Migration Status	-0.35	0.01
	Secondary School	-0.04	-0.01
	Clinical School	0.51*	-0.13**

Note: *significant at p < .05

** significant at p < .01

***significant at p < .001

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 October 2018 doi:10.20944/preprints201810.0455.v1
Peer-reviewed version available at Soc. Sci. 2019, 8, 46; doi:10.3390/socsci8020046

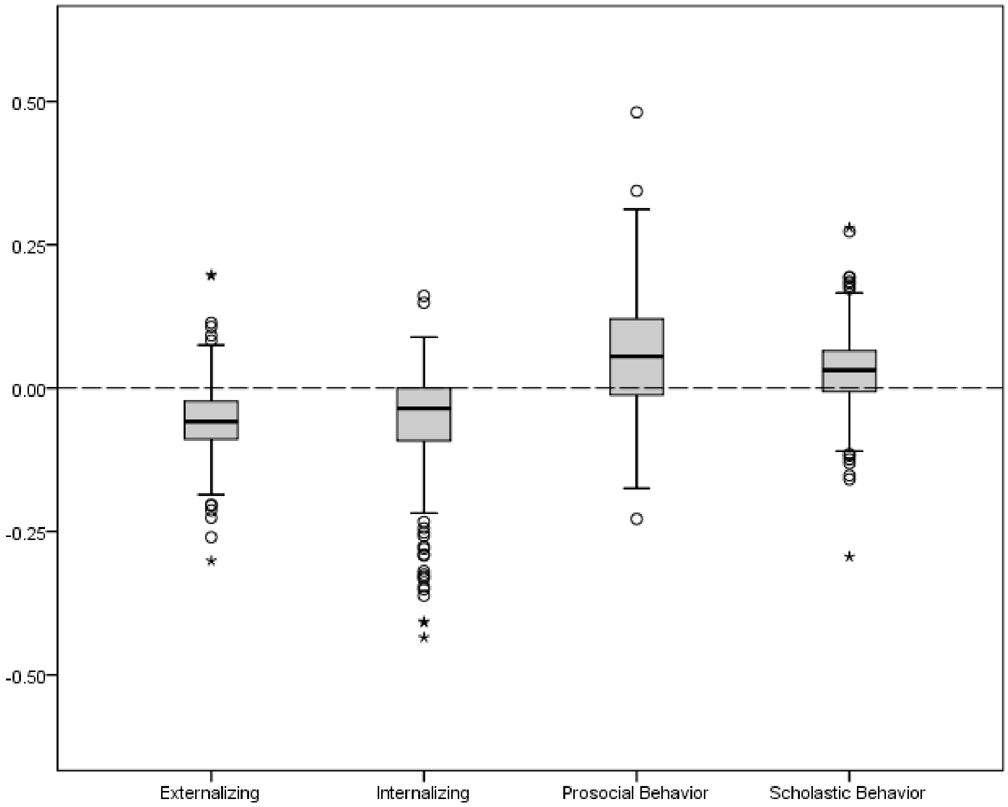

Figure A2. Individual unstandardized slopes for the latent growth models

References

1. Achenbach, Thomas M., and Craig S. Edelbrock. 1978. The classification of child psychopathology: A review and analysis of empirical efforts. *Psychological bulletin* 85 (6): pp. 1275–301. doi: 10.1037/0033-2909.85.6.1275.
2. Briesch, Amy M., Sandra M. Chafouleas, and T. Chris Riley-Tillman. 2010. Generalizability and Dependability of Behavior Assessment Methods to Estimate Academic Engagement: A Comparison of Systematic Direct Observation and Direct Behavior Rating. *School Psychology Review* 39: pp. 408–421.
3. Briesch, Amy M., Robert J. Volpe, and Tyler D. Ferguson. 2014. The influence of student characteristics on the dependability of behavioral observation data. *School psychology quarterly* 29 (2): 171–81. doi: 10.1037/spq0000042.
4. Casale, Gino, Thomas Hennemann, Robert J. Volpe, Amy M. Briesch, and Michael Grosche. 2015. Generalisierbarkeit und Zuverlässigkeit von Direkten Verhaltensbeurteilungen des Lern- und Arbeitsverhaltens in einer inklusiven Grundschulklasse [Generalizability and dependability of direct behavior ratings of academically engaged behavior in an inclusive classroom setting]. *Empirische Sonderpädagogik* (3): pp.258-268.
5. Center, David B., and John M. Callaway. 2017. Self-Reported Job Stress and Personality in Teachers of Students with Emotional or Behavioral Disorders. *Behavioral Disorders* 25 (1): 41–51. doi: 10.1177/019874299902500102.
6. Chafouleas, Sandra M. 2011. Direct behavior rating: A review of the issues and research in its development. *Education and Treatment of Children* 34 (4): pp. 575-591.
7. Costello, E. J., Sarah Mustillo, Alaattin Erkanli, Gordon Keeler, & Adrian Angold. 2003. Prevalence and development of psychiatric disorders in childhood and adolescence. *Archives of general psychiatry*, 60 (8), pp 837-844. doi: 10.1001/archpsyc.60.8.837
8. Chopin, Bruce. 1968. Item Bank using Sample-free Calibration, *Nature* 219, pp 870-872.
9. Christ, Theodore J., T. C. Riley-Tillman, and Sandra M. Chafouleas. 2009. Foundation for the Development and Use of Direct Behavior Rating (DBR) to Assess and Evaluate Student Behavior. *Assessment for Effective Intervention* 34 (4): pp. 201–13. doi: 10.1177/1534508409340390.
10. Cronbach, Lee J., Goldine C. Gleser, Harinder Nanda, and Nageswari Rajaratnam. 1972. The Dependability of Behavioral Measures. Theory of Generalizability of Scores and Profiles. New York: Jon Wiley & Sons.
11. Daniels, Brian, Robert J. Volpe, Amy M. Briesch, and Gregory A. Fabiano. 2014. Development of a problem-focused behavioral screener linked to evidence-based intervention. *School psychology quarterly* 29 (4): pp. 438–51. doi: 10.1037/spq0000100.
12. Deno, Stanley L. 2005. Problem solving assessment. In *Assessment for intervention: A problem-solving approach* R. Brown-Chidsey (Ed.), pp. 10-42, New York: Guilford Press.
13. DeVries, Jeffrey M., Stefan Voß, and Markus Gebhardt. 2018. Do learners with special education needs really feel included? Evidence from the Perception of Inclusion Questionnaire and Strengths and Difficulties Questionnaire. *Research in developmental disabilities* 83: pp. 28–36. doi: 10.1016/j.ridd.2018.07.007.
14. Dimitrov, Dimiter M. 2017. Testing for Factorial Invariance in the Context of Construct Validation. *Measurement and Evaluation in Counseling and Development* 43 (2): 121–49. doi: 10.1177/0748175610373459.
15. Durlak, Joseph. A., Roger. P. Weissberg, Allison. B. Dymnicki, Rebecca. D. Taylor, & Kriston B. Schellinger. 2011. The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child development* 82(1), pp 405-432. doi: 10.1111/j.1467-8624.2010.01564.x
16. Eklund, Katie, Tyler L. Renshaw, Erin Dowdy, Shane R. Jimerson, Shelly R. Hart, Camille N. Jones, and James Earhart. 2009. Early Identification of Behavioral and Emotional Problems in Youth: Universal Screening versus Teacher-Referral Identification. *The California School Psychologist 14 (1)*, pp. 89-95.
17. Fabiano, Gregory A. and Kellina Pyle. 2018. Best Practices in School Mental Health for Attention-Deficit/ Hyperactivity Disorder: A Framework for Intervention. School Mental Health. Online First, pp. 1-20.
18. Gebhardt, Markus, Jeffrey M. DeVries, Jana Jungjohann & Gino Casale. 2018. Questionnaire Monitoring Behavior in Schools (QMBS) DBR-MIS. Description of the scale „Questionnaire Monitoring Behavior in Schools" (QMBS) in English and German language. http://dx.doi.org/10.17877/DE290R-19139
19. Gebhardt, Markus, Jörg-Henrik Heine, Nina Zeuch, and Natalie Förster. 2015. Lernverlaufsdiagnostik im Mathematikunterricht der zweiten Klasse: Raschanalysen und Empfehlungen zur Adaptation eines Testverfahrens für den Einsatz in inklusiven Klassen. [Learning progress assessment in mathematic in second grade: Rasch analysis and recommendations for adaptation of a test instrument for inclusive

classrooms] *Empirische Sonderpädagogik* (7): pp. 206–22. http://www.psychologie-aktuell.com/fileadmin/download/esp/3-2015_20150904/ esp_3-2015_206-222.pdf

20. Good, Roland H., and G. Jefferson. 1998. Contemporary perspectives on curriculum-based measurement validity. In *Advanced applications of curriculum-based measurement*, ed. Mark R. Shinn, pp. 61–88. The Guilford school practitioner series. New York: Guilford Press.
21. Goodman, Robert. 1997. The strengths and difficulties questionnaire: A research note. *Journal of Child Psychology and Psychiatry* 38 (5), pp 581–586. https://doi.org/10.1111/j.1469-7610.1997.tb01545.x
22. Goodman, Robert. 2001. Psychometric properties of the Strengths and Difficulties Questionnaire. *Journal of the American Academy of Child and Adolescent Psychiatry* 40 (11), pp 1337–1345. http://dx.doi.org/10.1097/00004583-200111000-00015.
23. Goodman, Anna, Donna L. Lamping, and George B. Ploubidis. 2010. When to use broader internalising and externalising subscales instead of the hypothesised five subscales on the Strengths and Difficulties Questionnaire (SDQ): Data from British parents, teachers and children. *Journal of abnormal child psychology* 38 (8): pp. 1179–91. doi: 10.1007/s10802-010-9434-x.
24. Heine, Jörg H. 2014. pairwise: Rasch Model Parameters by Pairwise Algorithm. Computer software. Munich.
25. Heine, Jörg H., Markus Gebhardt, Susanne Schwab, Phillip Neumann, Julia Gorges, and Elke Wild. 2018. Testing psychometric properties of the CFT 1-R for students with special educational needs. *Psychological Test and Assessment Modeling* 60 (1), pp 3-27.
26. Heine, Jörg H., and Christian Tarnai. 2015. Pairwise Rasch model item parameter recovery under sparse data conditions. *Psychological Test and Assessment Modeling* 57 (1), 3-36.
27. Huber, Christian and Christian Rietz. 2015. Direct Behavior Rating (DBR) als Methode zur Verhaltensverlaufsdiagnostik in der Schule: Ein systematisches Review von Methodenstudien. *Empirische Sonderpädagogik*, 7(2), pp. 75–98.
28. Krull, Johanna, Jürgen Wilbert, and Thomas Hennemann. 2018. Does social exclusion by classmates lead to behaviour problems and learning difficulties or vice versa? A cross-lagged panel analysis. *European Journal of Special Needs Education* 33 (2): pp. 235–53. doi: 10.1080/08856257.2018.1424780.
29. Levine, Douglas W., Robert M. Kaplan, Daniel F. Kripke, Deborah J. Bowen, Michelle J. Naughton, and Sally A. Shumaker. 2003. Factor structure and measurement invariance of the Women's Health Initiative Insomnia Rating Scale. *Psychological Assessment* 15 (2): pp. 123–36. doi: 10.1037/1040-3590.15.2.123.
30. Linacre, John. M. .1994. Sample size and item calibration stability. *Rasch Measurement Transactions*, 7 (4), p. 328.
31. Linacre, John M. 2002. Optimizing rating scale category effectiveness. *J Appl Meas* 3.1. pp 85-106.
32. Moffitt, Terrie E., Avshalom Caspi, Honalee Harrington, and Barry J. Milne. 2002. Males on the life-course-persistent and adolescence-limited antisocial pathways: Follow-up at age 26 years. *Development and psychopathology* 14 (1): pp. 179–207.
33. Owens, Julie S., and Steven W. Evans. 2017. Progress Monitoring Change in Children's Social, Emotional, and Behavioral Functioning: Advancing the State of the Science. *Assessment for Effective Intervention* 43 (2): pp. 67–70. doi: 10.1177/1534508417737040.
34. R Core Team 2013. R: A language and environment for statistical computing.
35. Reinke, Wendy M., Keith C. Herman, Hanno Petras, and Nicholas S. Ialongo. 2008. Empirically derived subtypes of child academic and behavior problems: Co-occurrence and distal outcomes. *Journal of abnormal child psychology* 36 (5): pp. 759–7chop0. doi: 10.1007/s10802-007-9208-2.
36. Voss, Stefan, and Markus Gebhardt. 2017. Monitoring der sozial-emotionalen Situation von Grundschülerinnen und Grundschülern - Ist der SDQ ein geeignetes Verfahren? [Monitoring of the social emotional situation of elementary school students –Is the SDQ a suitable instrument?] *Empirische Sonderpädagogik*, (1), pp. 19-35. Retrieve under: http://www.psychologie-aktuell.com/fileadmin/download /esp/1-2017_20170810/esp_1-2017_19-35.pdf
37. Volpe, Robert J., and Amy M. Briesch. 2012. Generalizability and Dependability of Single-Item and Multiple-Item Direct Behavior Rating Scales for Engagement and Disruptive Behavior. School Psychology Review 41 (3), pp. 246–261.
38. Volpe, Robert. J., and Amy M. Briesch. 2015. Multi-item direct behavior ratings: Dependability of two levels of assessment specificity. *School Psychology Quarterly*, 30(3), pp 431-442. doi: 10.1037/spq0000115

39. Volpe, Robert J., Amy M. Briesch, and Sandra M. Chafouleas. 2010. Linking Screening for Emotional and Behavioral Problems to Problem-Solving Efforts: An Adaptive Model of Behavioral Assessment. *Assessment for Effective Intervention* 35 (4): pp. 240–44. doi: 10.1177/1534508410377194.
40. Volpe, Robert J., Amy M. Briesch, and Kenneth D. Gadow. 2011. The efficiency of behavior rating scales to assess inattentive-overactive and oppositional-defiant behaviors: Applying generalizability theory to streamline assessment. *Journal of school psychology* 49 (1): pp. 131–55. doi: 10.1016/j.jsp.2010.09.005.
41. Volpe, Robert J., Gino Casale, Changiz Mohiyeddini, Michael Grosche, Thomas Hennemann, Amy M. Briesch, and Brian Daniels. 2018. A universal behavioral screener linked to personalized classroom interventions: Psychometric characteristics in a large sample of German schoolchildren. *Journal of school psychology* 66: pp. 25–40. doi: 10.1016/j.jsp.2017.11.003.
42. Volpe, Robert J., James C. DiPerna, John M. Hintze, and Edward S. Shapiro. 2005. Observing students in classroom settings: A review of seven coding schemes. *School Psychology Review*, 34(4): pp. 454-74.
43. Warm, T.A. 1989 Weighted Likelihood Estimation of Ability in Item Response Theory. *Psychometrika* 54, 427-450.
44. Waschbusch Daniel A., Breaux Rosanna P. and Babinski Dara E. 2018 School-Based Interventions for Aggression and Defiance in Youth: A Framework for Evidence-Based Practice. *School Mental Health*, Online first, pp. 1-14Wright Benjamin D., Linacre John M. 1994. Reasonable mean-square fit values. *Rasch Measurement Transactions* 8:3 pp.370
45. Wright, Benjamin. D., and Geofferey. N. Masters. 1982. *Rating scale analysis*. Mesa Press: Chicago