KG-publicatie nr. 17

Measuring challenging student behavior. An overview of methodological properties and decisions

Huub Everaert

KG-17

H. Everaert
Measuring challenging student behavior. An overview of methodological properties and decisions.

Correspondentie over deze KG-publicatie kunt u sturen naar:
E-mail:huub.everaert@hu.nl

---

nr.1 Everaert, H.A. en J.C. van der Wolf (2005).
Behaviorally Challenging Students and Teacher Stress.

---

nr.2 Wolf, J.C. van der & J.M.F. Touw (2005).
Onderzoek naar zorg in het curriculum van de Theo Thijssen Academie.

---

nr.3 Doorn, E.C. van (2005).
De Gedragingenlijst voor Leraren (Tweede onderzoeksrapport).

---

nr.4 Wolf, J.C. van der (2005).
Probleemouders en de school: een onderontwikkeld terrein.

---

nr.5 Enthoven, M. (2005).
The contribution of the school environment to youths' resilience: A Dutch middle-adolescent perspective.

---

nr.6 Enthoven, M.; A.C. Bouwer; J.C. Van der Wolf & A. van Peet (2005).
Recognizing Resilience: Development and Validation of an Instrument to Recognize Resilience in Dutch Middle-Adolescents.

---

nr.7 Velderman, H & H.A. Everaert (2005).
Time-out or switch? (Paper presented at the ECER conference On 9 September 2005, University College Dublin).

---

nr.8 Touw, J.M.F., J.T.E. van Beukering & H.A. Everaert (2005).
Teachers' Personal Constructs on Problem Behavior (Paper presented at the annual meeting of the European Educational Research Association (EERA), Dublin, Ireland, September 7-10).

---

nr.9 Doorn , E.C. van (2005).
Levend leren: daar ga ik voor!

---

nr.10 J.T.E. van Beukering, J.M.F. Touw & H. Everaert (2005).
Teachers' personal constructs on problem behaviour: towards professional development
& Kos, P. (2005). Personal constructs on (problem) pupils: a teacher's view.
2 Papers presented at the International Practitioner Research Conference & Collaborative Action Research Network Conference (PRAR 2005), Utrecht, The Netherlands, November 4-6, 2005.

---

nr. 11 Everaert, H. & A. van Peet (2006). Kwalitatief en kwantitatief onderzoek.

---

nr. 12 Everaert, H.A. & J.C. van der Wolf (2006).
A Comparison of Stress and Burnout between Dutch General and Special Education Teachers. Paper presented at the annual meeting of the American Educational Research Association (AERA), San Francisco, USA, April 7-11, 2006.

---

nr. 13 Everaert, H.A. & J.C. van der Wolf (2006). Gender Perceptions of Challenging Student Behavior and Teacher Stress.

---

nr. 14 Peet, A.A.J. van (2006). Schaalconstructie.

---

nr. 15 Dellevoet, S.M.E., Beukering, J.T.E. van, Everaert, H.A. & Touw, J.M.F. (2006). Evaluatie van de methode Under Construction op Instituut Theo Thijssen.

---

nr. 16 Peet, A.A.J. van (2006). Q-sort. Een rangordening.

---

nr. 17 Everaert, H.A. (2007). Measuring challenging student behavior. An overview of methodological properties and decisions.

Inhoudsopgave

1 Inleiding

In de periode 2001-2006 is door Huub Everaert en Kees van der Wolf onderzoek verricht naar de manier waarop leraren naar kinderen met moeilijk gedrag kijken. In al deze projecten is de Nederlandse leraar gevraagd om het kind met de grootste gedragsproblemen in gedachten te nemen en daarover een aantal vragen te beantwoorden. In 2006 waren we zover om het door ons ontwikkelde meetinstrument ook buiten Nederland te gebruiken. In samenwerking met collega-onderzoekers uit China, Italië, Suriname, Rusland, Verenigde Staten en Zuid-Afrika hebben we in even zovele landen de door ons ontwikkelde vragenlijst om de interactie tussen leraar en meest problematische kind vast te leggen, afgenomen. In de lente van 2006 hebben in totaal 3500 leraren aan dit internationaal onderzoek deel genomen.

Momenteel zijn we bezig met het analyseren van de data en het schrijven van een boek waarin de onderzoekers uit de deelnemende landen verslag doen van hun bevindingen. De redactie van dit boek ligt in handen van Rob Roeser van Tufts University (Boston, VS) en de beide Nederlandse onderzoekers. Het boek zelf verschijnt naar alle waarschijnlijkheid in 2008.

Een belangrijke pijler en tevens hoofdstuk van dit boek is de ontwikkeling en kwaliteit van het gebruikte meetinstrument. Vanaf 2001 zijn Huub Everaert en Kees van der Wolf bezig geweest met het meten van de manier waarop de leraar naar het meest problematische kind in zijn eigen klas keek. In de 17e KG-publicatie van de kenniskring ‘Gedragsproblemen in de onderwijspraktijk’ is deze ontwikkeling stapsgewijs vastgelegd. Aangezien deze KG-publicatie uiteindelijk grotendeels in het boek zelf zal worden opgenomen, is de eigenlijke tekst van dit KG verder in het Engels.

2 Short overview of all research projects

In the Netherlands a first sample of teachers was drawn in 2001 in order to measure the interaction between challenging students and teachers. Five years later data concerning this relationship were collected in the USA, Russia, Suriname, South Africa, Italy, China (Hong Kong), and the Netherlands. Apart from broadening our local scope to an international context, we also continued working on the psychometric quality of assessing classroom transactions between teachers and challenging students. This chapter provides an overview of the most important methodological issues, decisions and underpinnings of the Student Questionnaire (SQ) carried out over the years 2001-2006.

The purpose of the SQ has been to develop a measuring instrument to map (1) different types of challenging student behavior as seen by teachers and (2) the associated nuisance or stress perceived by teachers. Viewed this way, types of challenging behavior are conceived as explicitly subjective views of teachers. This approach contrasts with the well-known, objective use of DSM-IV-RT criteria set by psychiatrist and trained psychologists. Our purpose is not in the least to disregard DSM-IV criteria or diagnoses. Given daily, normal activities of teachers and students in classrooms all over the world, we are looking for a non-clinical approach to map the mutual interaction between teachers and students. Coping with challenging students is considered an integral part of teachers' everyday work. We want to find out how cumbersome, tedious, and indeed stressful, this part of the job is.

Table 1. Overview Research Projects Challenging Students and Teacher Stress

	General sampling information			Part 2b					
Project	*School type*	*Sampling date*	*Sample size*	*Number of items*	*Theoretical inspiration*	*Incidence or hinder*	*Subject*	*Likert scale*	*Most important raised methodological issues with respect to part 2b*
A	Primary school	Autumn 2001	70	72 items	Brophy[a]	incidence	most challenging student	1-7	Frasing and wording of items; incidence of behavior of most challenging versus comparison student.
B1	Primary school	Autumn 2002	154	32 items	Brophy	incidence	most challenging student, comparison student, and students in general	1-7	Getting grip on the feasibility and usefulness of measuring challenging behavior by Likert scales. Q sort used as pivot.
B2	Primary school	Autumn 2002	122	60 items	Brophy	incidence	most challenging student and comparison student	-	
C1	Primary school	Autumn 2003	320	47 items, ITS	Greene, Abidin, & Kmetz[b]	hinder	most challenging student and comparison student	1-5	Replication study of ITS, developed by Greene, Abidin, & Kmetz (1997). Type of analysis: MGM
C2	Primary school and Special Educational Needs (SEN)	Autumn 2003	329	72 items; all 47 items of parallel project C1 included	Brophy/Greene, Abidin, & Kmetz	incidence and hinder	most challenging student	1-5	Rewording and selecting items; 22 items for further research selected by EFA.
C3	Teachers visiting a conference on students with behavioral problems	February, 2004	136	15 items, selected out the original ITS (C1)	Greene, Abidin, & Kmetz	incidence and hinder	most challenging student	0-4	Verifying magnitude of slopes between incidence and hinder of several items perrtaining to the same scale and anchoring Likert scales at 0
D	Primary school and SEN	Autumn 2004	868	22 items	selection of 22 revisited items	incidence and hinder	most challenging student	0-4	Confirmation of selected model based on project C2; 22 items loading on 6 factors.
E1	Teachers visiting conferences on students with behavioral problems	November, 2004	187	22 items of project D included	see above	incidence and hinder	most challenging student	0-4	Anchoring Likert scales at 0
E2		February, 2005	59		see above	incidence and hinder	most challenging student	0-4	Anchoring Likert scales at 0
F	Primary schools, SEN and secondary schools in seven countries	Spring 2006	3527	23 items / 1 item of project D split in two separate items	see above	incidence and hinder	most challenging student	0-4	International comparison of most challenging student; EFA and CFA on Dutch data.
G1	Secondary schools	Spring 2005	75	31 items / all 23 items of project E included.	see above	incidence and hinder	most challenging student	0-4	Repeated measures on incidence and hinder of most challenging student; i.e. (1) same student are scored by different teachers and (2) same teachers fill out questionnaire once a year.
G2	Secondary schools	Summer 2006	202		see above	incidence and hinder	most challenging student	0-4	

[a]Brophy, J. (1996). Teaching problem students. New York: The Guilford press. [b]Greene, R.W., Abidin, R.R., Kmetz, C. (1997). The Index of Teaching Stress: A measure of student-teacher compatibility. Journal of School Psychology, 35(3), 239-259.

In Table 1 an overview is given of all projects measuring the incidence of and annoyance caused by challenging behavior. As will be discussed below, the first research projects, A, B1 and B2, were based on the work of Jere Brophy (1996). In these projects questionnaire items were developed and tested describing various types of challenging behaviour. We have reported quite extensively in Dutch on the beginnings of our quest to measure challenging student behavior and the nuisance it causes (Everaert, 2003).

With respect to organizing the studies, collecting samples and analyzing all data, research project C has been the most complicated one. This is partly reflected in the division of the project in three branches (C1, C2 and C3). The whole project is concerned with the Index of Teaching Stress of Greene, Abidin, and Kmetz (1997). In addition to a full replication study of the work of Greene et al. (1997) and concomitant analysis on item level, we also replicated an analysis of Konold and Abidin (2004) on scale level.

In 2003 we detected some theoretical misconceptions or theoretical problems in the work of Greene, Abidin, and Kmetz (1997). These authors mention the relationship between incidence and stress, but they never, in fact, measure the incidence of challenging behavior as a separate topic. As will be more extensively discussed below, the ITS measures perceived stress of some behavioral topics (Part A) and relates it to stress reactions in classroom or to the relationship between teacher and student (Part B). In project C2, we decided to measure (1) the incidence of some behavioral topics, (2) the related annoyance, and (3) the stress reaction reported by the teacher. At the same time, the results of research projects A, B1 and B2, measuring challenging student behavior based on work of Brophy (1996), had turned out to be very promising. Projects A and B resulted in a large set of well defined behavioral items. In project C we combined all this information in partly overlapping research projects.

Major methodological implications of project C1 were discussed at length in an AERA-paper presented in San Diego (Everaert & Van der Wolf, 2004a). In more comprehensive format conclusions of project C1 have been published in the article *Stress in the student-teacher relationship in Dutch schools: A replication study of Greene, Abidin, & Kmetz's index of teaching stress (ITS)* (Everaert & Van der Wolf, 2006). The main conclusions of these Dutch and English publications will be reiterated in this chapter.

Project C resulted in the selection of 22 items to measure incidence of and annoyance caused by challenging behavior (see SQ appendix, part 2B). In 2004, same design and items were used in research project D. New data were also collected in 2006. The core of this chapter is devoted to the results of two exploratory factor analyses (EFA, projects C and D) and the results of one confirmatory factor analysis (CFA, project F).

Working through all our datasets we sometimes wondered which way to go. The goal was clearly marked, but as a researcher we just seemed to run into every practical and theoretical problem you can think of. Also, statistical solutions are never as clear-cut as presented in the applied-research-section of the excellent textbooks or peer reviewed journal articles available. What we want to stress here is that, whatever the outcomes of the analyses, applied statistics is about making decisions. Given substantive theory, applied statistics is about making decisions, selecting methods, evaluating underlying assumptions and performing a lot, really a lot of computations.

3 Research projects A and B

The first research projects (A, B1, and B2) were, as mentioned before, all inspired by the work of Jere Brophy (1996). They have been fully described in *Het meten van de meester* [Measuring teachers] (Everaert, 2003). Starting with the translation of the vignettes of Brophy (1996), 72 items were developed in the course of project A to describe twelve types of difficult student behavior along four categories. Each type of behavior - (1) low-achieving students, (2) failure syndrome students, (3) overly perfectionistic students, (4) underachieving students, (5) hostile-aggressive students, (6) passive-aggressive students, (7) defiant students, (8) hyperactive students, (9) distractible students, (10) immature students, (11) students rejected by their peers, and (12) shy/withdrawn students - was covered by 6 items measuring the incidence of challenging behavior. Items and scales were extensively discussed with a group of professional teachers.

In project B1 out of these 72 items 32 were selected for further research by maximizing the scored item differences between the most challenging student and a 'comparison' student from the same class. Project B2 focussed on measuring 60 of the original 72 items with data collected by Q-methodology as opposed to Likert scales. No differences in ranking were found between the 32 items both projects had in common (Everaert, 2003). Using 7-point Likert items proved to be a sound and useful device of collecting data on the incidence of challenging behavior.

4 Research project C: reviewing the ITS developed by Greene, Abidin, and Kmetz (1997)

In 2003 our orientation broadened from the work of Brophy (1996) to views and ideas brought forward by Greene, Abidin, and Kmetz (1997).[1] In order to combine both approaches two research projects were organized in autumn 2003. First of all, a replication study of Greene et al. (1997) was conducted. Although we had already gratefully borrowed their idea of focussing on a specific student while rating challenging behavior, in project C1 their approach was copied in detail. The Index of Teaching Stress (ITS) was developed to measure the stress a teacher experiences as a function of the transactions he or she has with a specific student. Originally, the ITS was designed as a parallel to the Parenting Stress Index (PSI), developed by Abidin (1986, 1995). Part A of ITS consists of five subtests to measure problematic student behavior as stressful or frustrating: (1) ADHD, (2) emotional lability, (3) anxiety/withdrawal, (4) low ability/learning disability, and (5) aggressive/conduct disorder). The four subtests underlying Part B explore the perceptions of the impact of the behavior on the teacher and the teaching process, the sense of efficacy and satisfaction in working with the student and the nature of the interactions with other adults involved with the student. In Part A teachers are asked to indicate for each selected student which typical expressions of problematic behavior that they consider to be stressful or frustrating are applicable to that particular student. They do this by rating 47 items on a 5-point Likert scale for each student assessed. In Part B (1) self-doubt/needs support, (2) loss of satisfaction from teaching, (3) disruption of the teaching process, and (4) frustration caused by working with parents are explored. For this purpose teachers rate another 43 statements (also on a 5-point Likert scale) which explore the impact of both the student's and the parent's behavior on the teacher and teaching process. The scores of the nine subtests of Parts A and B (90 items in total) are added up to assess the total stress a teacher is experiencing (Greene et al.; Konold & Abidin, 2004).

In our research project C1, the Index of Teaching Stress was translated into Dutch and teachers were asked to fill in the questionnaire twice, the first time with respect to the most behaviorally challenging student in the class room and the second time with respect to the seventh student on their class roster (i.e. 'comparison' student). A very attractive characteristic of ITS is the ability to show how specific aspects of student behavior may contribute to the (lack of) well-being of the teacher.

Multiple Group Method (MGM) was used to examine whether our data supported the factor structure found by Greene et al. (1997). The main conclusions of this replication study of ITS with respect to Part A are that 7 items out of 47 deserve closer attention (Everaert & Van der Wolf, 2006). In seven cases, the correlation between an item and other items congeneric to the same scale was lower than the correlation of the item and other factors, that is, factors the item was theoretically not supposed to measure. For both types of students assessed the ITS factor

[1] Over the years 2001-2003 numerous students of the Utrecht University of Applied Sciences participated in the research. We highly appreciate the enthusiasm and support of Simone Bakelaar, Brigit Bel, Theunis van den Berge, Edith Blom, Helen Blom, Sjoerd van Bommel, Cees Cobussen, Monique van Deursen, Gelske de Vries, Annette Dekkers, Menno van Es, Thirza Geerts-de Leeuw van Weenen, Ietje Gerbrandy, Thomas Heuschmid, Petra den Hollander, Mimi Kok, Dick Kooistra, Margot Koster-van de Staaij, Gea van der Meer, Ingrid Muurman, Leon Plomp, Ellen Posthumus, Bob van der Schaaf, Corine Schalij, Wim Smit, Inger Steinmetz, Rudolf ter Velde, Paul van der Velde, Carla Walstijn, Marieke Wolfsen, Lies Ykema, Grete van der Zaag-Tebbenhof in conducting part of the fieldwork.

structure was only partly reproduced in the Dutch sample. More important differences were found in case of scoring the 'comparison' student. Some teachers reported that although the items are valid for behaviorally challenging students, they are simply too extreme to describe the behavior of those comparison students. The second and partly related criticism teachers reported in the questionnaire had to do with the structure used to rate frequency versus stress of the response set of Part A. As teachers remarked, how can teachers rate distress if the student does not display the behavior described?

In the second research project of 2003 (project C2), items emanating from the work of Brophy (1996) and Greene et al. (1997) were combined in one questionnaire. Although all ninety items of Part A and B of Greene et al. were incorporated, it cannot be considered a replication study in the classical sense of research project C1. First of all, items concerning the behavior of the most challenging student were randomly mixed with items covering Brophy's vignettes. Secondly, all questions concerning the 'comparison' student were dropped. Thirdly, teachers were asked to rate the items the describing most challenging student both for incidence and for annoyance caused. To generate more statistical power respondents of projects C1 and C2 were merged and with a total sample of 652 subjects, we repeated an analysis performed by one of the cofounders of ITS (Konold & Abidin, 2004). To assess the correctness of the theoretical underpinnings of ITS an explorative factor analysis was conducted on the number and nature of the nine original *stress* scales (instead of ninety stress items). The theoretical properties of ITS predict that an explorative factor analysis should result in a clear-cut two factor solution. The first factor is composed of five constructs dealing with problematic student behavior and the second factor measures the overall level of distress experienced by the teacher as a function of self-perceptions and expectations in relation to the student. To compare our results with Konold and Abidin (2004), both maximum likelihood and principal axis factor analyses with varimax rotations were employed. Ironically, Dutch data outperformed the American data gathered by Konold and Abidin (2004). Although they did not publish the actual factor loadings, these authors state that "the Attention/Hyperactivity [ADHD] scale was found to exhibit large factor loadings on both factors. Moreover, the small difference between the two factor loadings tended to favor the Attention/Hyperactivity subtest as a measure of the teacher domain. Thus, from a practical scale interpretation perspective, the decision was made to treat the Attention/Hyperactivity scale as a separate domain" (Konold & Abidin, 2004, p.4). As said before, contrary to the American results, Dutch results were in concordance with the original theoretical model (Everaert & Van der Wolf, 2004b).

Project C2 had also resulted in data on the *incidence* of challenging behavior and a second round of EFA on scale level. Three different extraction methods followed by Varimax rotation were conducted, namely, Maximum Likelihood (ML), Principal Axis Factoring (PAF), and Principal Components Analysis (PCA). These analyses was by their very nature limited to the sampled respondents in project C2 (N = 305) and dealt with 5 incidence scales pertaining to Part A and 4 scales covering teacher stress of Part B of Greene et al. (1997). As expected, all scales of Part B ended up in one factor. However, scales measuring incidence of types of student behavior broke up in two different factors instead of one. On the one hand, scales tapping externalizing behavior like ADHD, emotional lability/low adaptability, and aggressive/conduct disorder banded together in one factor, while, on the other, there was a (strong) tendency for anxiety/withdrawal and low ability/learning disabled to end up in a separate factor (Everaert & Van der Wolf, 2004b). To be on the safe side, it may be concluded that this division in scales suggests a much more complicated relationship between incidence and nuisance experienced than was suggested in the work of Greene et al. Also, the remarks made on this topic by teachers in project C1 can be recalled here.

As a matter of fact, the developers of the ITS themselves draw attention to this in their original study.The ITS was developed under the assumption that the level of a teacher's distress regarding the specific behaviors of a given student is not merely a reflection of the frequency of the behaviors. This assumption emanates from the extensive cognitive/social learning literature […] positing that an individual's response to an event is a function of his or her affective and perceptual appraisal of the event and not merely the frequency of the event. (Greene et al., p. 241).

We endorse this point of view. However, in their effort to "present preliminary empirical evidence in this paper regarding the frequency versus perception issue as related to the ITS" (p. 242), they implicitly suggest that the linear relationship between incidence and annoyance is the same for all items part of the same scale. This point was explicitly covered in a small research project we conducted in February 2004 (research project C3).

February 2004, 136 primary school teachers (84% female and average years of teachers experience were 14.5 years) were sampled at a conference on students with behavioral problems. The conference took place in Amsterdam and was co-organized by the Utrecht University of Applied Sciences. Also note the semantic and substantive difference between a conference about students with behavioral problems and the topic of the SQ where teachers had to focus on the most challenging student. In project C3 fifteen items were selected for further study and five regression models were tested, each explaining perceived stress as a variable dependent on incidence. In every model the overall scale regression equation was compared to three separate item regression equations. With regard to emotional lability/low adaptabiliy (F (4, 118) = 3.23, p = .0148) and low ability/learning disabled (F (4, 119) = 3.93, p = .0049) the null hypothesis, stating the overall scale regression equation has the same intercept and slope as the three separate items making up this particular scale, was rejected in favour of three separate item based regression equations (Everaert & Van der Wolf, 2004a, 2006). That is, within both scales the relationship between perceived stress (dependent variable) and incidence (independent variable) cannot adequately be described by one common single regression line, suggesting a far more complicated pattern between incidence and perceived stress than envisaged by Greene et al. (1997). There is a striking parallel between the results of project C2 and C3. In both studies, items measuring incidence of and nuisance caused by more internalizing aspects of challenging behavior (viz., emotional lability/low adaptability and low ability/learning disabled) behave differently from items covering more expressive or energetic challenging students (viz., ADHD, anxiety/withdrawal, and aggressive/conduct disorder).

Summing up, important theoretical 'pillars' of ITS were neither confirmed in our research projects of 2003-2004 nor in other replication studies. On the basis of MGM analyses several items needed reconsideration (Everaert & Van der Wolf, 2004a, 2006). The description of challenging behavior was often formulated in such extreme wordings that it could not be used in case of 'comparison' students (see also Abidin & Robinson, 2002). EFA and regression analyses showed a much more complicated interaction between incidence and perceived stress/hinder (Everaert & Van der Wolf, 2004a, 2006). Confirmatory Factor Analysis (CFA) also cast some doubts on the scale Disruption of teaching process. In the theoretical underpinnings of Greene et al. (1997) this scale should load on part B, in our studies it turned out crossloading on Part A and loading on Part B (Everaert & Van der Wolf, 2004b). One of the cofounders of the ITS suggested dropping scale ADHD of Part A without theoretical substantial reasoning (Konold & Abidin, 2004). Despite all these comments and criticisms working with ITS has been very fruitful. The core idea of ITS

measuring teacher stress by focusing on the relationship between a teacher and the most behaviorally challenging student proved to be steady as a rock. Summer 2004 it became clear we had to plunge back in our item pool collected in the autumn of 2003, but the fruitful ITS-concept of focussing on the most challenging student was there to stay.

5 Research project C: exploring the incidence of challenging students

Up until now we still had not fully used all data gathered in the course of research project C. In describing the ITS replication study, we mentioned that 47 items of challenging behavior were also part of the second research project (C2) started autumn 2003. The main purpose of this project was to decipher what types of challenging behavior teachers are confronted with in their classrooms and how these relate to their daily workload. In this respect project C2 was the logical follow up of the earlier projects of 2001-2002 and Brophy's original vignettes continued to function as red thread in project C2. Earlier research projects A, B1, and B2 had resulted in a load of items to chart incidence of challenging behavior. Making a clear distinction between incidence and perceived stress associated with student behavior should double that load. Just incorporating ITS items in project C2 would result in such a huge number of items that no teacher would ever be willing to fill out such a questionnaire. Thus, in order to keep the project manageable, the decision was made to mix the input based on Brophy (1996) with that from Greene et al. (1997). It turned out that 27 items out of the original 47 developed by Greene et al. covered identical aspects of behavior to those defined by Brophy (1996) and could, accordingly, be used in the process of operationalizing Brophy's view on problem students. Apart from these 47 items, another 25 unique items originating from our previous Brophy-inspired research in 2001-2002 were added. In this way we succeeded to combine 47 (=20 + **27**) items delineating five scales as defined by Greene et al. with 52 (=25 + **27**) items operationalizing twelve vignettes of Brophy (1996) resulting all together in 72 (=20 + 25 + **27**) items.[2] This pool of items was rewritten by a team of educational experts of the Utrecht University of Applied Sciences and subsequently discussed, reviewed and elaborated at length in various settings with teachers themselves. Contrary to the view expressed by Greene et al., the decision was made to measure both incidence and perceived stress of the challenging behavioral aspect. In order to keep the printed outlay of the questionnaire within reasonable limits, the anchors of the Likert scales were rescaled from 1-7 to 1-5.

Explorative factor analysis was used to analyze this pool of 52 items measuring incidence of challenging behavior. It simply did not result in a perfect match. On a sample of 268 respondents, principal axis factoring (PAF) resulted in twelve factors extracting just 53% of the variance. To correct for measurement error, PAF was selected over principal component analysis (PCA) and the number of initial factors was determined by selecting those factors for which eigenvalues are greater than 1. For the sake of convenience and interpretability the decision was made to choose an orthogonal rotation. Bearing in mind the comorbidity of challenging behavior, selecting an oblique rotation would have more appropriate. However, unknown correlations between unknown factors at that time could easily have evolved in a unfounded process of trial and error. The first three varimax rotated factors showed seven or more items loading on one factor each. Especially with regard to factors at the bottom of the solution, several items loaded on two or more factors. Communalities (h^2) ranged from .27 to .75. In a way, it is just what was to be expected of a first exploratory factor analysis; there is nothing unusual in that. Theoretical interpretation of loadings in the factor structure matrix appeared to result in six, clear cut factors, covered by three to four items each. In order to eliminate double loading items and simultaneously keeping at least three items loading on one factor each, one new item had to be added to the original pool of 52 items. Out of this pool of 53 items, 21 items already picked out in the first round plus the newly added

[2] With the exception of hostile aggressive students (6 items), distractible students (5 items), and hyperactive students (5 items), all other operationalized vignettes of problem students (that is, failure syndrome students, low-achieving students, immature students, underachieving students, passive-aggressive students, overly perfectionistic students, shy/withdrawn students, defiant students, and students rejected by their peers) were covered by 4 items each.

one were selected for further analysis.[3] Searching a new factor solution that excluded the eliminated items is often recommended (Brown, 2006; Parker, Endler, & Bagby, 1993; Pett, Lackey, & Sullivan, 2003). The selected items were once more subjected to factor analysis under the same conditions (i.e., orthogonal PAF, eigenvalues > 1 with varimax rotation). The correlation input matrix of these selected 22 items is presented at the top op Table 2. We decided to present the input correlation matrices with item means and standard deviations of projects C, D, and F, so readers can repeat and broaden our analysis for themselves. The result of the principal axis factoring of data collected in the autumn of 2003, is shown in Table 3.

Table 2. Input Data (Correlations, Means and Standard Deviations) for Exploratory Factor Analysis (EFA) in Research Projects C and D and for the Initial Confirmatory Factor Analysis (CFA) in Research Project F

Research Project C (N = 284)

	F2 P2BQ01i	F2 P2BQ02i	F3 P2BQ03i	F5 P2BQ04i	F4 P2BQ05i	F6 P2BQ06i	F1 P2BQ07i	F5 P2BQ08i	F6 P2BQ09i	F6 P2BQ10i	F3 P2BQ11i	F3 P2BQ12i	F4 P2BQ13i	F4 P2BQ14i	F4 P2BQ15i	F5 P2BQ16i	F1 P2BQ17i	F4 P2BQ18i	F1 P2BQ19i	F2 P2BQ20i	F2 P2BQ21i	F3 P2BQ22i	F1 P2BQ23i
F2 P2BQ01i	--																						
F2 P2BQ02i	.460	--																					
F3 P2BQ03i	-.010	.032	--																				
F5 P2BQ04i	.028	.021	.254	--																			
F4 P2BQ05i	.126	.025	.185	.104	--																		
F6 P2BQ06i	.258	.077	.061	.162	.185	--																	
F1 P2BQ07i	.411	.252	-.037	.063	.107	.343	--																
F5 P2BQ08i	-.022	-.055	.296	.446	.155	.062	-.007	--															
F6 P2BQ09i	.290	.172	-.042	.082	.158	.596	.426	.036	--														
F6 P2BQ10i	.256	.070	.008	-.001	.239	.584	.439	.027	.555	--													
F3 P2BQ11i	.015	-.021	.544	.133	.110	.053	-.028	.199	-.022	-.072	--												
F3 P2BQ12i	-.056	-.040	.604	.189	.180	.053	-.098	.314	-.014	-.011	.530	--											
F4 P2BQ13i																							
F4 P2BQ14i	.151	.162	.068	.131	.383	.207	.142	.189	.201	.242	-.078	.057		--									
F4 P2BQ15i	.348	.195	.063	.008	.369	.183	.248	.119	.198	.236	-.069	.087		.350	--								
F5 P2BQ16i	.067	.055	.154	.285	.198	.008	.033	.468	.050	.077	-.016	.134		.320	.170	--							
F1 P2BQ17i	.329	.165	-.022	.126	.069	.313	.455	.055	.346	.397	-.063	-.111		.141	.282	.098	--						
F4 P2BQ18i	.086	.031	.071	.118	.454	.210	.196	.150	.168	.309	-.035	.022		.498	.417	.274	.210	--					
F1 P2BQ19i	.201	.081	-.030	.072	.098	.295	.551	.075	.247	.371	-.046	-.007		.182	.207	.157	.501	.192	--				
F2 P2BQ20i	.399	.427	.066	.073	.137	.351	.344	-.056	.277	.172	.088	.078		.053	.246	.027	.131	.027	.148	--			
F2 P2BQ21i	.470	.561	-.087	-.031	.053	.286	.284	-.032	.269	.252	-.130	-.112		.139	.315	.049	.237	.012	.107	.454	--		
F3 P2BQ22i	.026	-.006	.523	.113	.013	-.013	-.096	.312	-.042	.001	.430	.526		.000	.107	.048	-.052	-.048	-.046	-.016	.002	--	
F1 P2BQ23i	.252	.174	.002	.165	.054	.387	.681	.089	.352	.361	-.036	-.078		.162	.159	.133	.541	.143	.697	.217	.160	-.104	--
M	4.049	3.887	3.000	2.169	3.077	2.655	3.623	2.271	2.535	2.651	3.349	3.095		3.380	4.137	2.750	2.954	3.275	3.049	3.789	3.454	3.289	3.063
SD	1.028	1.184	1.319	1.053	1.364	1.356	1.172	1.204	1.284	1.392	1.138	1.250		1.222	1.019	1.200	1.314	1.251	1.342	1.301	1.270	1.316	1.322

[3] Except for the added item F4 P2BQ05i "… makes more of a fuss than others /[P2BQ13i] … cries more often. [combined]," hardly any difference was found between the factor analysis based on 53 items compared to the original factor analysis based on 52 items. The 'same' twelve factors explained 53% of variance (*N = 267)*. Communalities ranged from .25 to .75.

Research Project D (N = 725)

F2 P2BQ01i	--																					
F2 P2BQ02i	.612	--																				
F3 P2BQ03i	.202	.239	--																			
F5 P2BQ04i	.141	.137	.416	--																		
F4 P2BQ05i	.098	.126	.290	.241	--																	
F6 P2BQ06i	.203	.248	.203	.149	.190	--																
F1 P2BQ07i	.488	.368	.102	.070	.134	.365	--															
F5 P2BQ08i	.089	.071	.324	.523	.242	.119	.103	--														
F6 P2BQ09i	.242	.254	.198	.130	.107	.705	.331	.129	--													
F6 P2BQ10i	.227	.207	.152	.104	.206	.565	.401	.057	.494	--												
F3 P2BQ11i	.186	.235	.623	.299	.170	.256	.142	.291	.257	.160	--											
F3 P2BQ12i	.096	.132	.601	.378	.255	.169	.055	.435	.187	.120	.565	--										
F4 P2BQ13i																						
F4 P2BQ14i	.303	.322	.322	.306	.433	.323	.292	.292	.250	.297	.257	.282		--								
F4 P2BQ15i	.436	.383	.175	.137	.175	.160	.324	.128	.182	.192	.177	.118		.314	--							
F5 P2BQ16i	.085	.055	.235	.382	.172	.222	.124	.505	.179	.129	.259	.367		.265	.098	--						
F1 P2BQ17i	.284	.147	.128	.107	.168	.357	.441	.127	.280	.339	.136	.087		.271	.207	.176	--					
F4 P2BQ18i	.107	.110	.238	.276	.484	.292	.223	.301	.183	.271	.221	.272		.531	.199	.276	.279	--				
F1 P2BQ19i	.269	.161	.079	.079	.134	.306	.504	.177	.276	.315	.095	.090		.277	.246	.121	.406	.294	--			
F2 P2BQ20i	.476	.616	.302	.159	.154	.248	.323	.075	.323	.189	.307	.186		.254	.380	.095	.165	.142	.112	--		
F2 P2BQ21i	.386	.545	.136	.175	.110	.122	.198	.100	.172	.127	.159	.075		.210	.323	.063	.044	.061	.078	.478	--	
F3 P2BQ22i	.098	.061	.473	.326	.156	.143	.023	.379	.103	.113	.432	.529		.194	.115	.282	.052	.183	.008	.126	.100	--
F1 P2BQ23i	.351	.254	.075	.088	.130	.359	.649	.112	.312	.353	.078	.070		.270	.205	.098	.432	.206	.592	.226	.109	.099	--
M	3.026	2.846	1.586	.964	1.564	1.015	2.406	1.087	1.141	1.356	1.771	1.630		2.268	3.233	1.523	1.634	1.891	1.654	2.763	2.261	1.676	1.946
SD	0.991	1.201	1.322	1.123	1.421	1.243	1.224	1.256	1.169	1.376	1.263	1.264		1.327	0.967	1.179	1.381	1.344	1.284	1.262	1.323	1.476	1.315

Research Project F (N = 323)

F2 P2BQ01i	--	.115	.037	.055	.022	.023	**.242**	.052	.062	.084	.126	.007	**.177**	-.091	.086	-.018	.062	-.032	-.010	-.056	-.077	-.010	.052
F2 P2BQ02i	.594	--	.024	.027	-.122	-.043	.125	-.065	-.043	-.085	.143	-.069	.029	-.016	-.045	-.001	-.065	**-.154**	-.136	-.007	-.010	-.069	-.116
F3 P2BQ03i	.190	.211	--	.059	-.069	-.001	-.049	-.096	-.097	.003	.006	.017	.016	-.023	-.042	-.125	.042	-.100	-.027	-.017	-.078	.013	-.059
F5 P2BQ04i	.112	.097	.255	--	**.214**	.079	-.057	.001	-.030	-.025	-.029	.104	-.030	.113	-.008	-.030	.053	.121	-.076	-.022	.021	.077	-.110
F4 P2BQ05i	.233	.136	.057	.312	--	.025	-.065	.133	-.024	.133	.024	.070	-.020	.036	-.008	.108	.067	**.152**	-.071	-.094	.065	-.069	**-.147**
F6 P2BQ06i	.223	.201	.144	.186	.174	--	.042	.053	.028	-.029	.037	.050	-.055	.110	-.008	.041	-.002	.146	-.016	-.023	.001	.012	-.066
F1 P2BQ07i	.466	.398	.064	.070	.135	.342	--	-.028	.088	.108	.143	-.020	.107	-.089	.072	-.020	-.077	-.035	-.079	.123	.087	.000	.040
F5 P2BQ08i	.143	.046	.214	.420	.288	.222	.173	--	-.034	.050	-.054	.119	-.091	.023	-.073	.008	.107	**.195**	-.034	-.011	.003	-.028	.025
F6 P2BQ09i	.267	.207	.051	.080	.130	.564	.397	.139	--	-.005	.043	-.064	-.047	-.011	-.087	-.068	-.081	.047	-.033	.009	-.001	-.053	-.074
F6 P2BQ10i	.277	.150	.142	.078	.278	.474	.398	.213	.512	--	.106	.007	-.001	-.021	.044	-.105	.053	**.217**	.011	.011	.073	-.013	.054
F3 P2BQ11i	.274	.324	.575	.161	.146	.176	.252	.245	.186	.240	--	-.038	.106	.077	.058	.016	.108	.102	.021	.121	.087	.026	.040
F3 P2BQ12i	.162	.120	.609	.301	.196	.196	.094	.431	.086	.147	.534	--	.028	.038	-.064	.093	.086	.039	-.014	-.035	-.074	-.018	-.045
F4 P2BQ13i	.544	.477	.234	.140	.334	.205	.454	.180	.220	.249	.317	.248	--	-.081	.015	-.051	.081	-.106	-.019	.084	.056	-.068	-.011
F4 P2BQ14i	.169	.301	.132	.234	.287	.293	.156	.215	.178	.155	.227	.194	.353	--	.051	.124	.022	**.194**	.062	-.038	-.052	-.053	-.070
F4 P2BQ15i	.441	.389	.170	.157	.335	.243	.407	.189	.171	.286	.262	.148	.608	.471	--	-.012	.042	.007	-.013	-.018	-.052	-.096	-.054
F5 P2BQ16i	.058	.093	.134	.321	.239	.182	.148	.563	.077	.031	.267	.355	.175	.284	.207	--	.102	**.161**	.023	.030	.027	-.017	.022
F1 P2BQ17i	.240	.153	.132	.154	.226	.237	.329	.267	.165	.284	.194	.176	.356	.217	.308	.235	--	.125	.098	.024	-.015	-.021	-.051
F4 P2BQ18i	.164	.085	.017	.213	.341	.285	.150	.340	.189	.351	.215	.157	.221	.425	.324	.282	.273	--	.107	-.136	-.086	-.030	-.014
F1 P2BQ19i	.228	.154	.092	.059	.141	.304	.463	.179	.295	.319	.137	.107	.349	.322	.343	.201	.529	.303	--	-.037	-.034	-.017	.028
F2 P2BQ20i	.489	.658	.196	.058	.200	.254	.434	.116	.294	.278	.327	.179	.594	.323	.476	.136	.271	.137	.293	--	.032	-.060	-.027
F2 P2BQ21i	.423	.600	.118	.095	.335	.256	.372	.119	.261	.319	.276	.123	.524	.279	.401	.125	.211	.164	.269	.727	--	-.059	-.037
F3 P2BQ22i	.126	.098	.536	.251	.043	.140	.100	.247	.078	.111	.531	.507	.126	.084	.092	.213	.058	.074	.089	.129	.114	--	-.022
F1 P2BQ23i	.306	.193	.068	.034	.079	.275	.618	.252	.276	.382	.163	.083	.381	.207	.326	.213	.409	.195	.642	.326	.286	.091	--
M	2.920	2.511	1.740	1.260	2.146	1.523	2.498	.969	1.331	1.520	2.118	1.502	2.616	2.217	2.836	1.232	1.266	1.836	1.601	2.452	2.077	1.545	1.824
SD	1.137	1.298	1.381	1.234	1.425	1.343	1.254	1.190	1.263	1.338	1.365	1.303	1.456	1.352	1.261	1.179	1.287	1.347	1.398	1.434	1.420	1.498	1.389

Note. For replication purposes, numbers are rounded to three digits. In case of research project F input correlations are presented in the lower triangle. Upper triangle shows standardized residuals of initial CFA model. Ten largest standardized residuals are printed in bold.

Table 3. Factor Loadings, Communalities (h^2), Percents of Variance and Covariance for Principal Axis Factoring (PAF) and Varimax Rotation on Incidence of Behavioral Challenging Items (N = 284, Autumn 2003)

Item	F_1 [a]	F_3	F_2	F_4	F_6	F_5	Communalities (h^2) 6 Factors
F1 P2BQ23i ... breaks rules on purpose.	.86						.81
F1 P2BQ19i ... deliberately seeks conflict with adults.	.76						.61
F1 P2BQ07i ... undermines the rules.	.68		.30				.62
F1 P2BQ17i ... is belligerent towards me.	.56						.42
F3 P2BQ12i ... always finds the work difficult.		.77					.62
F3 P2BQ03i ... needs everything to be spelled out for him/her.		.76					.62
F3 P2BQ11i ... has trouble following instructions.		.71					.51
F3 P2BQ22i ... has obvious learning difficulties.		.63					.42
F2 P2BQ21i ... is much more active than the other students.			.75				.61
F2 P2BQ02i ... is unable to sit still.			.73				.54
F2 P2BQ01i ... distracts the other students.			.61				.46
F2 P2BQ20i ... leaves his/her seat more often than other students.			.57				.40
F4 P2BQ18i ... is hard to reassure whenever he/she is upset.				.74			.60
F4 P2BQ05i ... makes more of a fuss than others /P2BQ13i ....cries more often. [combined].				.59			.40
F4 P2BQ14i ... is overly sensitive to moods.				.59			.43
F4 P2BQ15i ... shows a strong reaction when something happens.			.32	.55			.43
F6 P2BQ06i ... is destructive.					.79		.71
F6 P2BQ09i ... damages other people's property.					.65		.54
F6 P2BQ10i ... is very aggressive; hits, kicks, bites.	.32			.28	.61		.56
F5 P2BQ08i ... hands in work giving remarks such as: 'it will be wrong anyway'.		.27				.74	.63
F5 P2BQ16i ... is not quite satisfied with end results.				.30		.57	.42
F5 P2BQ04i ... ascribes success to good luck.						.52	.32
Eigenvalue	2.44	2.24	2.13	1.86	1.71	1.30	
Percent of variance	11.10	10.17	9.67	8.44	7.79	5.93	
Cumulative percent of variance	11.10	21.27	30.95	39.38	47.17	53.10	
Percent of covariance	20.89	19.18	18.24	15.92	14.64	11.13	

Note. Items with a cut-off loading value of .25 are not shown in the table. Numbering of items is based upon the random order in the questionnaire to be administered Spring, 2006.

[a]F1 Against the grain (i), F2 Full of activity/Easily distractible (i), F3 Needs a lot of attention/Weak student (i), F4 Easily upset (i), F5 Failure syndrome/Excessively perfectionist (i), and F6 Aggressive/Hostile (i).

Principal axis factoring with varimax rotation was performed on 22 items from a sample of 284 elementary and special education teachers. The factorability of the correlation matrix judged by Keyser-Meyer-Olkin test (KMO = .80) and Bartlett's test of sphericity (X^2 = 2345.06, df = 231, p <.0000) are positive. The MSA statistics vary from .71 to .87. That is all above .70 and phrased in the words of Kaiser himself, the statistics vary from "middling" to "meritorious" (Kaiser, 1974, p. 35 in Pett et al., 2003, p. 78). Six factors were extracted. Loadings of variables on factors, communalities, and percentages of variance and covariance are shown in Table 3. The numbering of the extracted factors is based on the first PAF on 52 items. The item numbers reflect the randomized order in the questionnaire used in 2006 (see research project F).

With the exception of the fifth factor (Failure syndrome/Excessively perfectionist) every factor in Table 3 has at least two factor loadings > .60; grossly facilitating the naming of the factors. Given the content of the items at face value, factors interpretation and labelling were sensible conceptually and turned out to be rather straightforward. Readers are encouraged to verify this conclusion in Table 3. Factors F1 Against the grain (i) (loadings between .56 and .86), F3 Needs a lot of attention/Weak student (i) (loadings between .63 and .77), F2 Full of activity/Easily distractible (i) (loadings between .57 and .75), and F4 Easily upset (i) (loadings between .55 and .74) load on four items each. The last item F4 P2BQ15i "… shows a strong reaction when something happens" can be judged as somewhat doubtful. Communality of this item is .43 and made up by two factor loadings of respectively .32 and .55. Cronbach's alphas were used to evaluate the factors' internal consistency. When item P2BQ05i was deleted from F4 Easily upset (i) the internal consistency dropped from .73 to .70, whereas excluding item P2BQ05i from the other items loading on F2 Full of activity/Easily distractible (i) lead to a higher Cronbach's alpha (from .76 with 5 items to .77 with 4 items). This all justified the conclusion to treat P2BQ05i as loading on F4.

Item F6 P2BQ10i "… is very aggressive; hits, kicks, bites" with a cut-off of loading value of .25, loads on three externalizing factors (F1, F2, and F6) and, judged by its content, the item is most closely related to F6 conceptually. Both other congeneric items (F6 P2BQ06i and F6 P2BQ09i) also capture aggressive behavior. With a cut-off of .25 six items load on F4 Easily upset (i). The same holds for F2 Full of activity/Easily distractible (i). Internal consistency varied from .84 (F1 Against the grain (i)) to .73 (F4 Easily upset (i)). Cronbach's alpha of F5 Failure syndrome/Excessively perfectionist (i) was rather low with .67.[4] A comment we made on the theoretical underpinning of ITS concerned the absence of a clear distinction between incidence and perceived stress of specific behavioral aspects. For the sake of completeness a principal axis factoring with varimax rotation was also performed on 22 items concerning perceived stress. (*N* = 272). This resulted in an identical factor structure, extracting over 60% of the variance.[5]

[4] Correlations among (clusters of) factors are discussed at the end of this chapter. See also Table 6.
[5] Result are available on request.

6 Research project D

The number of cases compared to factors is a much debated topic. Most popular is the rule of thumb of 10 cases for each item (Garson, n.d.). Regardless the subject-to-variables ratio Gorsuch (1983) advises 200 or more cases. Both criteria were met in the last stage of project C. That is, in the second factor analysis the ratio of subject-to-cases is just over 10. However, bearing in mind we selected 22 items out of a pool of 53 items, the real subject-to-cases is well below 10 (i.e., around 5). The questionnaire was put to the field once more and the autumn of 2004 725 teachers answered all 22 items on the most challenging student they could think of (response rate of 35%). Teachers, 78% of whom were female, reported an average class size of 20 students. Their average teaching experience was 14 years. Participants were mainly recruited in the Dutch provinces of Noord-Holland, Zuid-Holland, Zeeland, Utrecht, Gelderland and Noord-Brabant.[6] Prior to the analysis, the age of selected students was cut off at 13 years and variables were examined with regard to missing values. Data cleaning resulted in some loss of respondents. From the original 868 sampled participants 725 fully completed questionnaires could be used for further analysis. Because of the psychometric nature of the present studies only cases who completed all 22 SQ items were included in the analyses concerning projects C and D.

Research project D differs in two important ways from research project C. First of all, teachers were asked to score just 22 items instead of 47 (project C1) or 72 (project C2). In the world of the psychometric qualities of measuring instruments that is considered quite a difference. Secondly, the Likert scales were changed from 1-5 to 0-4. The decision to change the anchors was based on teachers reporting difficulties in rating a score of 1 in cases where the student did not exhibit any sign of the specific behavior. We had already pre-tested this change in research project C3 (Everaert & Van der Wolf, 2004a, 2006). Exploratory factor analysis is, of course, essentially an exploratory procedure and, given the role of sampling error, results of an initial EFA should be interpreted cautiously and, if possible, cross-validated using independent data sets (Brown, 2006). Both Kaiser-Meyer-Olkin measure of sampling adequacy of .86 and Bartlett's test of sphericity (χ^2= 6339.00, df = 231, p <.0000) indicated excellent factorability of the correlation matrix. MSA statistics varied from .81 to .93 and supported this conclusion. The same EFA criteria were applied as before, that is, eigenvalues >1, orthogonal solution followed by varimax rotation. The input correlation matrix of project D is presented in Table 2, results of the additional EFA on the new data set are shown in Table 4.

[6] Several students of Utrecht University of Applied Sciences participated in collecting the data: Arjanneke Brandsma, Sabine Bax, Menno van Es, Petra den Hollander, Frits van Hout, Gea Hoving, Gerbert Sipman, Lindy Slingerland, Albert Sluiter, Ingrid Muurman, Gerda Pool en Wil Vlam. We appreciate their efforts in sampling the respondents.

Table 4. Factor Loadings, Communalities (h^2), Percents of Variance and Covariance for Principal Axis Factoring (PAF) and Varimax Rotation on Incidence of Behavioral Challenging Items (*N* = 725, Autumn 2004)

							Communalities (h^2)
Item	F_2 [a]	F_1	F_3	F_6	F_4	F_5	**6 Factors**
F2 P2BQ02i ... is unable to sit still.	.82						.71
F2 P2BQ20i ... leaves his/her seat more often than other students.	.69						.55
F2 P2BQ01i ... distracts the other students.	.65	.35					.56
F2 P2BQ21i ... is much more active than the other students.	.64						.43
F4 P2BQ15i ... shows a strong reaction when something happens.	.46						.31
F1 P2BQ23i ... breaks rules on purpose.		.77					.65
F1 P2BQ07i ... undermines the rules.	.32	.72					.65
F1 P2BQ19i ... deliberately seeks conflict with adults.		.67					.50
F1 P2BQ17i ... is belligerent towards me.		.51					.35
F3 P2BQ03i ... needs everything to be spelled out for him/her.			.77				.67
F3 P2BQ12i ... always finds the work difficult.			.71			.29	.62
F3 P2BQ11i ... has trouble following instructions.			.70				.57
F3 P2BQ22i ... has obvious learning difficulties.			.56			.28	.40
F6 P2BQ06i ... is destructive.				.83			.80
F6 P2BQ09i ... damages other people's property.				.74			.65
F6 P2BQ10i ... is very aggressive; hits, kicks, bites.		.33		.52			.44
F4 P2BQ18i ... is hard to reassure whenever he/she is upset.					.70		.60
F4 P2BQ05i ... makes more of a fuss than others /P2BQ13i ...cries more often. [combined].					.60		.41
F4 P2BQ14i ... is overly sensitive to moods.	.26				.58		.53
F5 P2BQ08i ... hands in work giving remarks such as: 'it will be wrong anyway'.			.26			.79	.72
F5 P2BQ16i ... is not quite satisfied with end results.						.54	.39
F5 P2BQ04i ... ascribes success to good luck.			.31			.51	.42
Eigenvalue	2.55	2.33	2.28	1.77	1.50	1.49	
Percent of variance	11.59	10.61	10.38	8.05	6.82	6.77	
Cumulative percent of variance	11.59	22.20	32.58	40.63	47.45	54.22	
Percent of covariance	21.39	19.55	19.13	14.85	12.58	12.50	

Note. Items with a cut-off loading value of .25 are not shown in the table. Numbering of items is based upon the random order in the questionnaire to be administered Spring, 2006.

[a]F1 Against the grain (i), F2 Full of activity/Easily distractible (i), F3 Needs a lot of attention/Weak student (i), F4 Easily upset (i), F5 Failure syndrome/Excessively perfectionist (i), and F6 Aggressive/Hostile (i).

The similarity between both EFAs in Tables 3 and 4 is striking. Research project D resulted in the very same factor structure explaining 54% of variance. All items but one (F4 P2BQ15i) loaded on their original factor. Proportions of item variance explained by the factor structure (h^2) oscillated between .71 and .31. As expected, F4 P2BQ15i is the item with the lowest communality. Evaluating the sums of squared loadings for an item in an orthogonal factor matrix should be related to the extent which the item plays a role in the interpretation of the factor. This means that, given our theoretical underpinnings of item content, even low communalities could be interpreted as contributing meaningfully to their respective factor. To sum it all up, given the relatively low number of items per factor and the more or less balanced values of the communalities within their congeneric factors, the EFA solution of research project D may be judged as favourable or positive.

Over two independent samples a replicable and interpretable factor structure was extracted from the 22 item SQ. However, the cross-loading tendency of item F4 P2BQ15i deserves a closer look. In 2003 the item loaded on F4 Easily upset (i) (.55) and F2 Full of activity/Easily distractible (i) (.32) (see Table 3). While this tendency to cross-load is gone in project D, the overall solution deteriorated. Over two independent samples, item F4 P2BQ15i changed from loading on F4 (.55 in project C) to loading on F2 (.46 in project D). Reconsidering our theoretical approach of dealing with both externalizing (F1, F2, and F6) and internalizing (F2, F3, and F5) challenging student behavior, there seemed to be some overlap in factors F2 and F4. They were simply not as clear cut or separate form each other as we wanted them to be, but alas, border regions are often fluid by nature and change over history. In 2006 we decided to go ahead with all these items with just one exception.

7 Research project F

Confident that the SQ was an appropriate instrument for mapping teachers' views on the most challenging student, we set out to broaden our scope by crossing the Dutch border. The substantive and theoretical results of the Dutch part of this international project will be presented in chapter [4], here we will keep focussing on the methodological qualities of the SQ. In the spring of 2006 388 teachers participated in research project F. These respondents filled out 23 instead of 22 items on the incidence of challenging student behavior. When compared to research projects C and D the combined item F4 P2BQ05i "... makes more of a fuss than others/[P2BQ13i] ... cries more often" was split into two separate items (i.e., F4 P2BQ05i "... makes more of a fuss than others," and P2BQ13i "...cries more often"). As an insight admittedly too late, but in hindsight rather obvious as well: an item should cover only one issue at a time. Long and complicated items tend to be less valid. At the same time, the Dutch translation of F4 P2BQ13i "...cries more often" was changed. In projects C and D the verb to cry in item F4 P2BQ13i was translated in 'shedding tears.' In project F the translation focussed on yelling or shouting aloud.

The 23 item battery was administered to 359 (special) primary school teachers. Of this group 323 subjects completed all incidence items and were included in the analyses. Almost 79% percent of these subjects were female. Their teaching experience was on average 17.4 years (*SD* = 11.3). More boys (79%) than girls (21%) were selected as the most challenging pupils. The average age of the children selected by the teachers was just under nine years old (*M* = 8.7, *SD* = 2.1). Reported class size averaged about 21 students. As in all research projects discussed so far, students of the Utrecht University of Applied Sciences collected the data.[7] Questionnaires were left at the teacher rooms or handed out at staff and teacher meetings. An introductory letter stated purpose and importance of the research. Participants were given the option of returning the questionnaire in a prepaid envelope to the university. In several cases, surveys were sealed in an envelope and collected at the school itself. As in former research projects, anonymity was guaranteed. Using convenience samples makes it difficult to give accurate response rates. Hardly any different results could be traced to the fieldworkers that collected the data.

After several research projects and numerous analyses, we now have a clear sense of the number of factors that exist in the data, and of the items that are related to the various factors. That is the moment when CFA comes into view. EFA is in essence data-driven. The objective of EFA is to evaluate the minimum number of interpretable factors to explain the correlations among a set of items, or indicators, as they are usually called. In CFA numbers of factors and patterns of item-factor loadings are a priori specified. A fundamental strength of CFA approaches is their ability to deal explicitly with measurement errors (Brown, 2006). CFA is typically conducted after one or more EFAs to foster development and refinement of the measurement model. CFA provides answers that help establish the convergent and discriminant validity of our theoretical constructs. A last advantage of CFA over EFA is the availability of several explicit goodness-of-fit criteria. In CFA, the specified factor solution is evaluated in terms of how well the sample correlation or covariance matrix of the measured items is reproduced.
Several CFA models were fit using EQS 6.1 (Bentler, 2005). All analyses were based on the raw data matrices, a necessary requirement in the robust analysis of categorical data. Although

[7] The assistance of Sabine Bax, Hellen Blom, Annette Dekkers, Frits van Hout, Gerrit de Peuter, Mimi Poll, Ellen Posthumus, and Bob van der Schaaf with doing fieldwork in project F is gratefully appreciated.

imputation of missing values based upon the EM algorithm may be considered standard by now in SEM (Enders, 2006, 2007), given the psychometric nature of project F, only subjects who filled out all 23 incidence items were included in the analysis. In order to correct for data characteristics that did not perfectly meet assumptions underlying normal theory estimators, the Satorra-Bentler scaled χ^2 and derived statistics were employed. This strategy is often recommended in case of non-normality and ordered categorical data (Finney & DiStefano, 2006). Evaluation of model fit was based on multiple criteria that reflected statistical, theoretical and practical consideration. Four different robust fit indices will be reported: the S-B χ^2 statistic, the comparative fit index (CFI), the normed fit index (NFI), the non-normed fit index (NNFI), and the root mean-square error of approximation (RMSEA) with the 90%-confidence interval. In general, the target values for the selected fit indices CFI, NFI and NNFI should be ≥.95, while a RSMEA ≤ .05 indicates also good model fit. In evaluating correlation residuals (also labelled standardized residuals in EQS output), figures >|1.0| indicate big difference between observed and predicted covariances (Kline, 2005).

The initial CFA model hypothesized a priori that: (a) responses to the incidence of challenging student behavior could be explained by six factors; (b) each of the 23 items would have a non-zero loading on the factor it was designed to measure, and zero-loadings on all other factors; (c) the six factors would be correlated; and (d) error/uniqueness terms for the item would be uncorrelated. That is, factor F4 Easily upset (i) loaded on 5 items, 3 factors (F1 Against the grain (i), F2 Full of activity/Easily distractible (i), and F3 Needs a lot of attention/Week student (i)) loaded on 4 items each and 2 factors (F5 Failure syndrome/Excessively perfectionist (i) and F6 Aggressive/Hostile (i)) were specified to load on 3 items.

At first glance and by generally applied standards, the fit of the initial CFA model should be considered unsatisfactorily (S-B χ^2 (215) = 660.841, p < .0000, CFI = .869, NFI=.812, NNFI=.846, RMSEA = .075 with 90% CI between .067 and .082, and about 84% of the standardized residuals are between -0.1 and 1.0). However, given specification differences by definition between EFA and CFA models in general, model fit may be judged as surprisingly good. Especially when taking into account that, when EFA is used as precursor to CFA, oblique solutions are more likely to generalize to CFA than orthogonal solutions (Brown, 2006).[8] After all, about 85% of covariance in the data set is explained by the oblique model as indicated by the CFI.

[8] The orthogonal CFA model was considerably worse (S-B χ^2 (230) = 1057.485, p < .0000, CFI = .719, NFI=.669, NNFI=.691, RMSEA = .106 with 90% CI between .099 and .112.). The chi-square difference for the orthogonal and oblique 23-item model was highly significant. However, strictly speaking, the Satorra-Bentler correction is not suited for comparing nested models.

Table 5. Maximum Likelihood Robust Parameter Estimates for Initial CFA Model (*N* = 323, Spring 2006)

Parameter	Unstanderdized	SE	Standerdized
	Factor loadings		
F1 → P2BQ07i	1.000[a]	--	.715
F1 → P2BQ17i	.815	.089	.568
F1 → P2BQ19i	1.183	.096	.758
F1 → P2BQ23i	1.253	.093	.809
F2 → P2BQ01i	1.000[a]	--	.627
F2 → P2BQ02i	1.394	.120	.765
F2 → P2BQ20i	1.753	.167	.870
F2 → P2BQ21i	1.592	.156	.799
F3 → P2BQ03i	1.000[a]	--	.768
F3 → P2BQ11i	.954	.068	.741
F3 → P2BQ12i	.949	.066	.772
F3 → P2BQ22i	.961	.070	.681
F4 → P2BQ05i	1.000[a]	--	.452
F4 → P2BQ13i	1.771	.205	.783
F4 → P2BQ14i	1.164	.163	.554
F4 → P2BQ15i	1.485	.166	.758
F4 → P2BQ18i	.875	.130	.418
F5 → P2BQ04i	1.000[a]	--	.515
F5 → P2BQ08i	1.525	.200	.814
F5 → P2BQ16i	1.266	.167	.682
F6 → P2BQ06i	1.000[a]	--	.722
F6 → P2BQ09i	.967	.091	.742
F6 → P2BQ10i	.961	.089	.696

Table 5 continued

Parameter	Unstanderdized	SE	Standerdized
	Measurement error variances		
$E_{P2BQ01i}$	.785	.069	.607
$E_{P2BQ02i}$	.700	.071	.415
$E_{P2BQ03i}$	.783	.114	.411
$E_{P2BQ04i}$	1.119	.100	.735
$E_{P2BQ05i}$	1.617	.106	.796
$E_{P2BQ06i}$	.864	.121	.479
$E_{P2BQ07i}$	.770	.078	.489
$E_{P2BQ08i}$	.478	.098	.337
$E_{P2BQ09i}$	.717	.091	.449
$E_{P2BQ10i}$	.922	.100	.515
$E_{P2BQ11i}$	.839	.087	.450
$E_{P2BQ12i}$	.686	.100	.404
$E_{P2BQ13i}$	.820	.087	.387
$E_{P2BQ14i}$	1.268	.102	.693
$E_{P2BQ15i}$	.677	.081	.426
$E_{P2BQ16i}$	.743	.116	.535
$E_{P2BQ17i}$	1.122	.108	.677
$E_{P2BQ18i}$	1.497	.107	.825
$E_{P2BQ19i}$	.831	.091	.425
$E_{P2BQ20i}$	.498	.074	.242
$E_{P2BQ21i}$	.730	.080	.362
$E_{P2BQ22i}$	1.204	.118	.537
$E_{P2BQ23i}$	.667	.091	.346

Variances and covariances of factors[b]

	F1	F2	F3	F4	F5	F6
	.804 (.108)	.507 (.100)	1.124 (.130)	.414 (.095)	.403 (.093)	.940 (.133)
F1 Against the grain (i)	--					
F2 Full of activity/Easily distractible (i)	.319 (.066)	--				
F3 Needs a lot of attention/Weak student (i)	.196 (.065)	.241 (.059)	--			
F4 Easily upset (i)	.357 (.063)	.343 (.066)	.248 (.055)	--		
F5 Failure syndrome/Excessively perfectionist (i)	.197 (.047)	.081 (.032)	.334 (.059)	.173 (.043)	--	
F6 Aggressive/Hostile (i)	.507 (.074)	.305 (.063)	.268 (.073)	.286 (.059)	.177 (.050)	--

Standerdized variances of factors

	F1	F2	F3	F4	F5	F6
F1 Against the grain (i)	1.000					
F2 Full of activity/Easily distractible (i)	.500	1.000				
F3 Needs a lot of attention/Weak student (i)	.206	.319	1.000			
F4 Easily upset (i)	.619	.748	.364	1.000		
F5 Failure syndrome/Excessively perfectionist (i)	.345	.179	.496	.424	1.000	
F6 Aggressive/Hostile (i)	.583	.442	.261	.459	.287	1.000

[a]Not tested for statistical significance; $p < .05$ for all other unstanderized estimates. [b]Robust standard errors between ().

Secondly, in order to detect strain in the initial oblique model solution, the largest standardized residuals were studied further. Of the 10 largest standardized residuals item F4 P2BQ18i showed up 5 times. Disconcertingly, the initial model did not explain the correlation between F4 P2BQ18i and F4 P2BQ05i (.341) and between F4 P2BQ18i and F4 P2BQ14i (.425). Item correlations of research project F are given at the bottom of Table 2 (the lower triangle). As can be seen in Table 5, standardized factor loading explained just .189 (=.418*.452) of the correlation between F4 P2BQ18i and F4 P2BQ05i. By the same reasoning the explained model correlation of .232 (=.418*.554) between F4 P2BQ18i and F4 P2BQ14i is not good enough. It seems fair to conclude that factor loadings of F4 P2BQ05i (.452), F4 P2BQ14i (.554) and F4 P2BQ18i (.418) are underestimated in the model. Also, R-square of F4 P2BQ18i is with a score of .175 well below the minimum standard criteria of .200 as given by Brown (2006). Simply splitting the original item F4 P2BQ05i used in projects C and D into F4 P2BQ05i and F4 P2BQ13i for project F, resulted in an extra item loading on F4. The input correlation matrix showed high correlations between F4 P2BQ13i and congeneric items F4 P2BQ05i (.334), F4 P2BQ14i (.353), and F4 P2BQ15i (.608), but not between items F4 P2BQ13i and F4 P2BQ18i (.221). Because of this, the standardized factor loading P2BQ13i on F4 is - compared to the other congeneric item loadings- probably too high (.783). Summing up, there must be some misspecification in factor F4.

Thirdly, the largest standardized residual is between F1 P2BQ07i and F2 P2BQ01i (.242); two items not supposed to load on the same factor. Total correlation between both items according to the specified model is just about .224 (=.715*.500*627) and that is far below the item correlation of .466 as presented in Table 2. The items have something in common that cannot be adequately explained by the product of two standardized loadings times the factor correlation between F1 and F2. The problem is exemplified in the phrasing of F1 P2BQ07i "... undermines the rules." Students can undermine rules in a variety of ways. The item does not describe typical behavior, it seems to be more of a general qualification of the student. Item P2BQ07i is worded too vaguely or generally, as may also be concluded from a closer inspection of the input correlation matrix *prior to* all this statistical reasoning. Twelve out of 22 possible F4 P2BQ07i item correlations are above. 300. Item F4 P2BQ07i indicates a student is just challenging, without specifying what kind of behavior the student exhibits according to the teacher.

Fourthly, the specified model cannot account for the correlation between F5 P2BQ04i and F4 P2BQ05i (.312), resulting in the third largest standardized residual (.214). The total product of both standardized loadings (.452*.515) and the standardized covariance between F4, F5 (.424) is just .096. Reflection on item content explains why both items correlate reasonably well. Item F5 P2BQ04i focuses on external attribution and students who exhibit this kind of behavior also experience fewer grip on their situation. That may result in making a fuss and that is exactly where item F4 P2BQ05i comes in. Unfortunately, the initially specified CFA model does not account for this correlation.

Modification indices help to address strain in the model. The Lagrange multiplier indicates freeing three parameters at least (F4 P2BQ13i loading on F2, (χ^2 =35.03, df =1, p <.0000); F1 P2BQ07i loading on F2 (χ^2=26.77, df =1, p <.0000); and F3 P2BQ012i loading on F5 (χ^2=25.39, df =1, p <.0000). Misspecification issues of P2BQ13i and P2BQ07i have already been dealt with. Specifying a cross loading of F3 P2BQ012i on F5 may be (partly) explained by taking into account the labels of the respective factors (F3 Needs a lot of attention/Weak student (i) and F5 Failure syndrome/Excessively perfectionist (i)). More specific, a Failure syndrome student/Excessively perfectionist student (F5) will probably often find the work difficult. Item F3 P2BQ012i reads " …

always finds the work difficult." Counter to the Lagrange multiplier test, the Wald does indicate constraining item loadings to zero. Invoking the Wald test at this stage in model specification would not be appropriate.[9]

On the basis of the strain detected in the original CFA model, it is probably not too difficult to find a well fitting CFA six factor model. A few minor adaptations - deletion of malfunctioning items P2BQ07i and P2BQ18i, allowing a couple of items to cross-load and incorporating some correlated errors - will probably do the job. Although referring to multilevel modelling, the words of Bickel (2007, p. 206) however echo too loudly to proceed: "After all, how many times can a regression model be respecified before we acknowledge that we are making too much of the peculiarities of one data set." At this stage of the research process, that is not the issue, anyway. The question at issue is whether we are satisfied with the structure found in the data to proceed with analyses based upon our original idea of specifying six latent factors divided over internalizing and externalizing challenging student behavior. To answer this question, we will first have to address the discriminant validity of the SQ. In Table 6 Cronbach's alphas of and correlations between factor scores of research projects C, D, and F are presented.

[9] Another way of addressing strain in the model is by exploratory factor analysis within the framework of confirmatory factor analysis (E/CFA) (Brown, White, Forsyth, & Barlow, 2004; Brown, 2006). In an eloquent treatise, Brown (2006) gives the exact model specifications as follows: (a) fixing factor variances to unity, (b) freely estimating factors covariances, (c) selecting an anchor item for each factor whose cross loadings on all other factors are fixed to zero, and (d) loadings of nonanchor items are freely estimated on each factor. Problem in following this strategy was selecting the anchor items. This should be based upon exploratory factor analysis under the same conditions as in projects C and D. In this case the solution resulted in a five factor solution. Indeed, original factor F4 dissolves in F2 and F5. That is, items supposed to load on F4 turn out to load on F2 (F4 P2BQ13i and F4 P2BQ15i) and F5 (F4 P2BQ05i, F4 P2BQ14i, and F4 P2BQ18i).

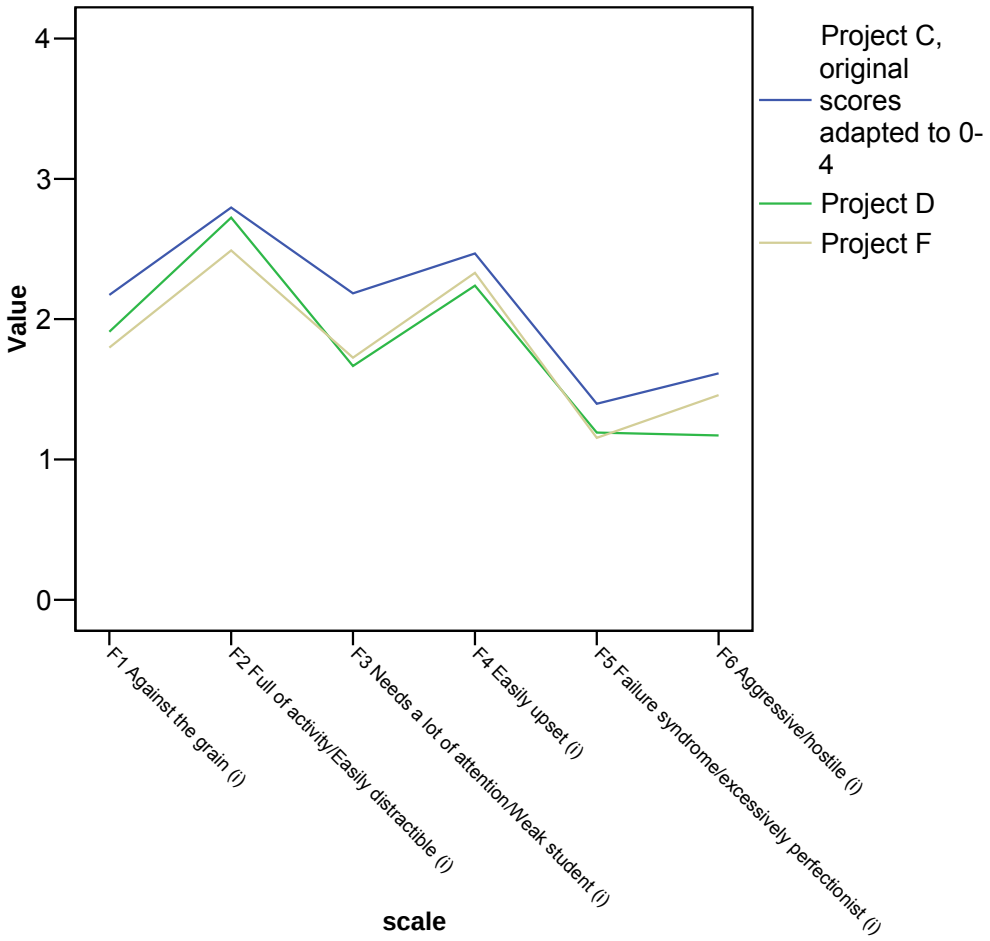
Table 6. Factor Correlations and Internal Consistency of Factors in Research Projects C, D, and F

	F1	F2	F3	F4	F5	F6	Cronbach's alpha
Research Project C (*N* = 284)							
F1 Against the grain (i)	--						.841
F2 Full of activity/Easily distractible (i)	.332	--					.772
F3 Needs a lot of attention/Weak student (i)	-.080	-.016	--				.816
F4 Easily upset (i)	.254	.210	.082	--			.734
F5 Failure syndrome/Excessively perfectionist (i)	.139	.015	.288	.286	--		.667
F6 Aggressive/Hostile (i)	.509	.347	-.004	.334	.083	--	.804
M	3.173	3.795	3.183	3.467	2.397	2.614	
SD	1.061	0.925	1.009	0.910	0.894	1.140	
Reseach Project D (*N* = 725)							
F1 Against the grain (i)	--						.800
F2 Full of activity/Easily distractible (i)	.335	--					.807
F3 Needs a lot of attention/Weak student (i)	.128	.253	--				.819
F4 Easily upset (i)	.390	.350	.368	--			.699
F5 Failure syndrome/Excessively perfectionist (i)	.183	.159	.512	.400	--		.728
F6 Aggressive/Hostile (i)	.496	.309	.245	.361	.194	--	.805
M	1.910	2.724	1.666	2.239	1.191	1.171	
SD	1.030	0.956	1.074	0.925	0.955	1.074	
Reseach Project F (*N* = 323)							
F1 Against the grain (i)	--						.801
F2 Full of activity/Easily distractible (i)	.434	--					.849
F3 Needs a lot of attention/Weak student (i)	.182	.272	--				.828
F4 Easily upset (i)	.477	.549	.271	--			.743
F5 Failure syndrome/Excessively perfectionist (i)	.263	.152	.396	.415	--		.696
F6 Aggressive/Hostile (i)	.470	.365	.211	.393	.208	--	.761
M	1.797	2.490	1.726	2.330	1.154	1.458	
SD	1.055	1.101	1.127	0.962	0.947	1.082	

Note. Correlations among factors of research projects C, D and F are based upon summing scores of items 'by hand,' an approach advocated by Pedhazur and Schmelkin (1991), instead on factor scores automatically generated by the respective EFA analysis. For Research project F, correlations among factors based upon the intial CFA model are also presented in Table 5. In research project C Likert-items varied from 1 to 5. In projects D and F items scores varied from 0 to 4.

In all projects there is a moderate to fairly-strong correlation between F1 Against the grain (i), F2 Full of activity/Easily distractible (i), and F6 Aggressive/Hostile (i). The correlation between F1 and F6 has been always strong (.509 in project C, .496 in project D, and .470 in project F). The correlations between F1 and F2 (and concomitantly between F6 and F2) is also in line with the theoretical underpinnings (.332 in project C, .335 in project D, and .434 in project F) of three factors delineating externalizing student behavior. With regard to internalizing challenging student behavior, correlations between internalizing factors are not as clear. In all research projects correlations between F3 Needs a lot of attention/Weak student (i) and F5 Failure syndrome/Excessively perfectionist (i) are in accordance with our main view (.288 in project C, .512 in project D, and .396 in project F).The same holds more or less for correlations between F4 Easily upset (i) and F5 Failure syndrome/Excessively perfectionist (i). It seems justified to conclude that there are indeed two clusters of different types of challenging student behavior: internalizing versus externalizing. However, the correlations between F4 and various externalizing factors (e.g. .334 with F6 in project C, .390 with F1 in project D, and .477 and .549 with F1 and F6 respectively in project F) are less convincing. As can be seen in Table 6, the homogeneity of scales expressed by Cronbach's alphas is in general above the criterion of > .70 set by Nunnally (1978). In Figure 1 mean scores of all factors used in projects C, D, and F are presented. Mean scores in project C are rescaled to 0 to 4 instead of the original 1 to 5. The uniformity over the different projects is striking.

Figure 1. Means of six Incidence Scales by Project.



8 Conclusion

Theoretical and substantive views put forward by Brophy (1996) and Greene et al. (1997) were incorporated in our diverse research projects. Our research benefited from their pioneering work on the behavior of challenging students. The latent structure in all research projects is fairly constant over time. That is, over three independent samples within a time span of about 2.5 years, a replicable and interpretable factor structure was extracted from the data sets. There is considerable coherency and consistency in the covariance structures of incidence of challenging student behavior as viewed by teachers. Also, mean structure is very similar over the years and it is about time to proceed with analyzing the international data sets and forget about the psychometric qualities. As a matter of fact, in Table 6 we already had left the world of latent factors.

Unfortunately, not just good psychometric qualities are reproduced over the years, items loading on F4 even seemed to deteriorate over the years. We are fully aware that there is still some psychometric work to be done. In future research projects we will reconsider the phrasing and exact wording of the used items, especially as regards F4. We have deliberately published items that did not live up to our expectations (viz., P2BQ07i and P2BQ18i) instead of deleting them from the research projects C and D on basis of results sampled in 2006. We have given some insight in the decisions we have taken in the whole process and some of the decisions simply did not work out very well. Perhaps not everyone will acknowledge this explicitly, but it is an integral part of applied psychometric statistics. After all, who is to benefit from dressing up a beautiful, shining latent world while the researchers know by themselves it is covered with non-random spatters of mud. Working with convenience samples can be rather inconvenient, as well.

9 References

Abidin, R.R. (1986). Parenting Stress Index (2nd ed.). Charlottesville, VA: Pediatric Psychology Press.

Abidin, R.R. (1995). Parenting Stress Index: Test manual (3rd ed.). Osessa, Fl: Psychological Assessment Resources.

Abidin, R.R. & Robinson, L.L. (2002). Stress, biases, or professionalism: What drives teachers' referral judgments of students with challenging behaviors. Journal of Emotional and Behavioral disorders, 10(4), 204-212.

Bentler, P.M. (2005). EQS 6 structural equations program manual. Encino, CA: Multivariate Software, Inc.

Bickel, R. (2007). *Multilevel analysis for applied research: It's just regression!* New York: The Guilford Press.

Brophy, J. (1996). *Teaching problem students*. New York: The Guilford press.

Brown, T.A., White, K.S., Forsyth, J.P., & Barlow, D.H. (2004). The structure of perceived emotional control: Psychometric properties of a revised Anxiety Control Questionnaire, *Behavior Therapy, 35*, 75-99.

Brown, T.A. (2006). Confirmatory factor analysis for applied research. New York, NJ: The Guilford Press.

Enders, C.K. (2006). Analyzing structural equation models with missing data. In G. Hancock & R.O. Mueller (Eds)., *Structural Equation Modeling: A Second Course* (pp. 313-342). Greenwich, Connecticut: Information Age Publishing.

Enders, C.K. (2007). Analysis of missing data. AERA professional development. Chicago, Ill April 11.

Everaert, H.A. (2003). Het meten van de meester. [Measuring Teacher Stress]. In K van der Wolf (Ed.). *Het hoofd van de meester* (pp. 33-61). Utrecht: Uitgeverij Agiel.

Everaert, H.A. & Van der Wolf, K.C. (2004a). Stress in the Student-Teacher Relationship in Dutch Schools: a replication study of Greene, Abidin, & Kmetz's Index of Teaching Stress-Instrument (ITS). Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA, April 12-16.

Everaert, H.A. & Van der Wolf, K.C (2004b). An Empirical Look at Underlying Theoretical Constructs of the Index of Teaching Stress (ITS). Paper presented at 25th international STAR Conference, Amsterdam, The Netherlands, July 7-10.

Everaert, H.A. & Wolf, K.C. van der (2006). Stress in the student-teacher relationship in Dutch schools: A replication study of Greene, Abidin & Kmetz's index of teaching stress (ITS). In R. Lambert & C. McCarthy (Eds.), Understanding teacher stress in an age of accountability (pp. 121-143). Greenwich, Connecticut: Information Age Publishing Inc.

Finney, S.J. & Distefano, C. (2006). Nonnormal and categorical data in structural equation. In G. Hancock & R.O. Mueller (Eds)., *Structural Equation Modeling: A Second Course* (pp. 269-314). Greenwich, Connecticut: Information Age Publishing.

Garson, G.D. (n.d.). Factor Anlysis, from *Statnotes: Topics in Multivariate Analysis.* Retrieved 05/18/2007 from http://www2.chass.ncsu.edu/garson/pa765/statnote.htm.

Gorsuch, R.L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Greene, R.W., Abidin, R.R., & Kmetz, C. (1997). The Index of Teaching Stress: A measure of student-teacher compatibility. Journal of School Psychology, 35 (3), 239-259.

Kline, R.B. (2005). *Principals and practice of structural equation modeling* (2nd ed.). New York: The Guilford Press.

Konold, T.R. & Abidin, R.R. (2004). A Look at the Factor Structure of the Index of Teaching Stress. Paper presented at the annual meeting of the American Educational Research. Association, San Diego, CA, April 12-16.

Nunnally, J.C. (1978). *Psychometric Theory*. New York: McGraw-Hill.

Parker, J.D.A., Endler, N.S., & Bagby, R.M. (1993). If it changes, it might be unstable: Examining the factor structure of the Ways of Coping Questionnaire. *Psychological assessment, 5*, 361-368.

Pedhazur, E.J. & Schmelkin, L.P. (1991). *Measurement, design, and analysis: An integrated approach.* Hillsdale, NJ: Lawrence Erlbaum.