



Humanwissenschaftliche Fakultät


Fabian Kirsch | Robert Busching | Helena Rohlf | Barbara Krahé

Using behavioral observation for the longitudinal study of anger regulation in middle childhood

Suggested citation referring to the original publication:
Applied Developmental Science (2017)
DOI <http://dx.doi.org/10.1080/10888691.2017.1325325>
ISSN (print) 1088-8691
ISSN (online) 1532-480X

Postprint archived at the Institutional Repository of the Potsdam University in:
Postprints der Universität Potsdam
Humanwissenschaftliche Reihe ; 461
ISSN 1866-8364
<http://nbn-resolving.de/urn:nbn:de:kobv:517-opus4-412557>

Using behavioral observation for the longitudinal study of anger regulation in middle childhood

Fabian Kirsch , Robert Busching, Helena Rohlf, and Barbara Krahe



University of Potsdam

ABSTRACT

Assessing anger regulation via self-reports is fraught with problems, especially among children. Behavioral observation provides an ecologically valid alternative for measuring anger regulation. The present study uses data from two waves of a longitudinal study to present a behavioral observation approach for measuring anger regulation in middle childhood. At T1, 599 children from Germany (6–10 years old) were observed during an anger eliciting task, and the use of anger regulation strategies was coded. At T2, 3 years later, the observation was repeated with an age-appropriate version of the same task. Partial metric measurement invariance over time demonstrated the structural equivalence of the two versions. Maladaptive anger regulation between the two time points showed moderate stability. Validity was established by showing correlations with aggressive behavior, peer problems, and conduct problems (concurrent and predictive criterion validity). The study presents an ecologically valid and economic approach to assessing anger regulation strategies *in situ*.

In everyday life, humans are constantly faced with different emotional states of positive and negative valence (Trampe, Quoidbach, & Taquet, 2015). As one of the basic emotions (Ekman & Friesen, 1971), anger has received special attention, not least because it has been associated with aggressive behavior in many studies (e.g., Wittmann, Arce, & Santisteban, 2008). In childhood in particular, anger is a very common and intense emotion (von Salisch, 2000). Children report greater difficulties (Waters & Thompson, 2014) and lower self-efficacy (Zeman & Shipman, 1997) in dealing with anger compared to other negatively valenced emotional states, such as sadness. To avoid the social and behavioral problems associated with unfiltered anger expression, children need to learn to deal with their anger in an appropriate way. That is, they have to learn the adaptive regulation of the emotional state of anger. There is a broad range of problems associated with maladaptive anger regulation, including aggressive behavior (e.g., Helmsen & Petermann, 2010), peer problems (e.g., von Salisch, Zeman, Luepschen, & Kanevski, 2014), and conduct problems (e.g., Morris, Silk, Steinberg, Terranova, & Kithakye, 2010).

Given the critical role of maladaptive anger regulation as a potential risk factor for behavioral and peer problems, studying the development of anger regulation on the basis of longitudinal designs is an important task. In the present research, we propose and validate a behavioral observation method that lends itself to the longitudinal study of anger regulation in middle childhood by exposing children to an anger-eliciting task at successive points in time and observing their regulation strategies. This methodological approach has the advantage of yielding information about a child's anger regulation in a real-life situation. At the same time, it faces the challenge of developing conceptually equivalent, but age-adapted versions of the anger-eliciting task. This is the basis for comparing anger regulation strategies over time. In earlier work, we assessed children's anger regulation strategies through observation and related them to aggressive behavior and peer problems cross-sectionally (Rohlf & Krahe, 2015). Based on this work, we designed an age-adapted task that was used to study the development of anger regulation in relation to developmental problems at a second assessment 3 years later. Specifically, we considered the role of maladaptive anger regulation as a predictor of aggressive behavior, peer problems, and conduct problems.

CONTACT Fabian Kirsch  fakirsch@uni-potsdam.de  Department of Psychology, University of Potsdam, Karl-Liebknecht-Str. 24-25, 14476 Potsdam, Germany.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hads.

© 2017 Fabian Kirsch, Robert Busching, Helena Rohlf, and Barbara Krahe

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

Published with license by Taylor & Francis

Measuring anger regulation in middle childhood

Past research on anger regulation in middle childhood has mostly used self-reports to measure children's anger regulation skills (Aldao, Nolen-Hoeksema, & Schweizer, 2010), but this procedure often involves difficulties (Underwood, 1997). Self-reported anger regulation may be biased as children may not be aware of their actual emotions or the way in which they try to regulate them. In actual anger eliciting situations, the emerging emotions may influence information processing (Berkowitz, 2012) and, consequently, the use of regulation strategies. Thus, self-reports may reflect children's knowledge of regulation strategies rather than their actual behavior. As a result, self-reports of the use of potential anger regulation strategies in hypothetical situations are unlikely to reflect the use of anger regulation strategies in real life. For example, in a study conducted by Parker et al. (2001), 6- to 11-year-old children generated fewer strategies in real-life anger eliciting situations than they mentioned for comparable hypothetical situations.

Given these limitations of self-reports, there is a need for more ecologically valid instruments that assess anger regulation in situations where anger is actually experienced. In order to address this task, Rohlf and Krahe (2015) developed and validated a behavioral observation method for children in the age group of 6 to 10 years. Anger was induced by having the children work on a virtually unsolvable dexterity task (i.e., a tower-building task). Anger regulation was assessed by means of a coding scheme that categorized children's specific regulation strategies, including visual and verbal focus on frustrating stimuli, venting the anger, resignation, and solution-oriented behavior. Strategies were classified as adaptive or maladaptive with respect to the prevention or promotion of negative interpersonal outcomes, such as aggressive behavior and social rejection, as explained in detail in Rohlf and Krahe (2015).

The *in-situ* method provides ecologically valid information about children's anger regulation strategies and avoids several of the problems of self-reports (Rohlf & Krahe, 2015; Kirsch, Rohlf, & Krahe, 2015). However, it poses challenges for the use in longitudinal studies, in particular with regard to the comparability of the anger-eliciting task. One way to address this problem is to use the exact same task (Wohlwill, 1973). However, presenting the same task repeatedly could lead to higher reactivity (including possible memory and training effects), which threatens the internal validity of the assessment (Schaie & Hofer, 2001). For instance, children who are exposed to exactly the same anger-arousing situation for the second time may use more effective strategies

to deal with their emotions as they are familiar and experienced with this setting. Additionally, using the same measure at different ages may not be appropriate because children become more cognitively and motorically skilled (e.g., Bartolotta & Shulman, 2010). As they get older, children may understand that the tower-building task is in fact impossible to complete, and the nature of the task may no longer be age-appropriate as they have outgrown playing with bricks. On the other hand, introducing a novel task can be potentially problematic as differences between the tasks may undermine the comparability of findings over time.

Thus, establishing the equivalence of anger-eliciting tasks used in the longitudinal study of anger regulation through behavioral observation is a methodological challenge. To maximize equivalence, the nature of the task should remain similar over time and require similar skills yet take into account maturational changes in behavior and interest.

The present study

The present study aimed to further develop the behavioral observation method of anger regulation presented by Rohlf and Krahe (2015) for use in longitudinal research studies. A major goal of our study was the development of a new observational measure to examine anger regulation longitudinally in middle childhood. The anger-eliciting task was selected to be appropriate for both boys and girls, that is without stereotypical gender preferences or connotations. This was deemed important to be able to evaluate potential gender differences irrespective of the characteristics of the specific anger-eliciting task. A few studies have revealed gender differences in the regulation of anger (e.g., Underwood, Coie, & Herbsman, 1992) and in the evaluation of the adaptiveness of specific anger regulation strategies (e.g., Waters & Thompson, 2014). In addition, we sought to demonstrate, in an exemplary fashion, the challenges and advantages of designing equivalent measures to study developmental processes over time.

We designed an anger-eliciting task that was conceptually equivalent to the tower-building task reported in Rohlf and Krahe (2015), but adapted to the children's improved cognitive and motor skills 3 years after the initial assessment. We provided support for the equivalence of both assessments and for the validity of our new, adapted task. In addition, we determined the stability of anger regulation over the course of 3 years. The structural equivalence of both tasks was established by assessing longitudinal measurement invariance between the observational measures at Time 1 (T1) and Time 2 (T2). For comparing regression slopes longitudinally,

it is essential that the latent factors for both time points have the same unit of measurement (Chen, 2007). Thus, we aimed to establish at least metric invariance.

A high correlation between the measures at T1 and T2 indicates construct stability, that is, consistency of anger regulation strategies across time. As emotion regulation is assumed to be a stable individual difference variable in middle childhood (Cole, Michel, & Teti, 1994) we expected a substantial correlation between our two measures.

To assess validity, we examined the relations between the T1 and T2 assessments of maladaptive anger regulation and aggressive behavior, peer problems, and conduct problems (criterion validity). These constructs were correlated with anger regulation in past research. For example, aggressive children used more maladaptive anger regulation strategies than did non-aggressive children, including focusing on frustrating stimuli and venting the anger (Helmsen & Petermann, 2010). Another study found cross-sectional and longitudinal relations between anger dysregulation (venting) and externalizing behavior (Morris et al., 2010). Regarding the peer context, anger regulation was identified as an important predictor of having reciprocal friendships (von Salisch et al., 2014).

The following hypotheses were examined in our study: (1) The new anger-eliciting task developed at T2 is conceptually equivalent to the task used at T1, as reflected in metric invariance between the two tasks. (2) Anger regulation observed at T1 is substantially correlated with anger regulation observed at T2 based on an age-adapted version of the anger-eliciting task (construct stability). (3) Anger regulation observed at T1 shows prospective associations and anger regulation at T2 shows concurrent associations with aggressive behavior, peer problems, and conduct problems (criterion validity).

Based on conceptual considerations as well as past empirical research (e.g., Morris et al., 2010), we assumed the proposed relations to hold for both boys and girls, and we expected no age differences. To address these assumptions, we tested the invariance of our measures across gender and age groups, and examined gender and age group differences in the proposed associations between our measures of anger regulation and the validation constructs.

Method

Participants and procedure

The sample consisted of 599 children from Germany (50.8% girls) who took part in the T1 assessment of a larger longitudinal study on intrapersonal

developmental risk factors in childhood and adolescence (see Appendix A for a more detailed description). They were between 6 and 10 years old at T1 ($M = 8.12$ years, $SD = 0.92$). Of these, 554 children (50.2% girls) took part in the T2 measurement about 3 years later, which corresponds to a high retention rate of 92.5%.¹ The time interval between T1 and T2 was on average 2.77 years ($SD = 0.19$). Thus, the children were between 9 and 13 years old at T2 ($M = 10.81$ years, $SD = 0.90$). To avoid a reduction in sample size, all 599 T1 participants were included in the study, and missing data were handled using multiple imputation, as described in the following section.

The sample was recruited from 33 community elementary schools representing a variety of rural and urban areas, and socio-economic backgrounds. In total, 33.9% of the mothers and 36.1% of the fathers reported holding a university degree, 22.9% and 13.6% had university entrance qualification, 41.6% and 48.9% had a vocational-level qualification, and 1.6% and 1.4% had no or a low level of school qualification. Because the study was conducted in a part of Germany where the ethnic diversity is low (i.e., mostly Caucasian), we did not explicitly ask about ethnic background. However, we asked the parents which language they primarily spoke with their children. The vast majority of the parents (94%) reported speaking exclusively German with their children, and only 0.4% reported exclusively speaking a foreign language.

At both waves, data collection took place at the participants' school and was conducted by trained project staff. The materials and procedure were approved by the Ethics Committee of the authors' university as well as the Ministry of Education, Youth and Sport of the Federal State of Brandenburg, Germany, where the study was conducted. Active written consent was obtained from the parents, and the children provided assent before the start of the data collection.

Anger-eliciting tasks

To observe children's use of anger regulation strategies, two similar yet age-adapted anger-eliciting tasks were designed for T1 and T2. In both tasks, participants were given a virtually impossible dexterity task. At T1, the task involved building a tower with 10 wooden blocks on the basis of a photo placed in front of the child. The tower collapsed every time, since two of these blocks were slightly rounded on one side. The T1 task

¹The data are part of a study that included three waves, in which the behavioral observation measure was employed at T1 and T3. Only data from these two waves are used in the present study. To avoid confusion, we refer to the two data waves as T1 and T2 for the purposes of the current analysis.

is described in detail in Rohlf and Krahe (2015). The task required the motor skills to balance the blocks to equilibrium. The new task developed for use 3 years later at T2 tapped into the same ability, but it was modified to be age-appropriate. The new task involved stacking seven dice to form a tower. The dice were of different sides, including two four-sided dice (tetrahedrons). Because of these two tetrahedrons, the dice tower task was practically impossible to accomplish (tetrahedrons only have one contact surface, which makes balancing the dice very difficult). A (photo-shopped) picture of a possible dice tower was placed in front of the children, but they were told that it just showed an example and they were free to stack the dice as they liked. The pictures used in the two tasks are presented in Appendix B. At both T1 and T2, there was a time restriction of 2 minutes and 40 seconds, determined in a pilot study as an appropriate time window for the display of anger regulation strategies. An hourglass (T1) or an electronic timer (T2) was placed on the table, so that the children could see the time running out. In addition, the children were told that they would get one of three presented gifts if they managed to complete the task in the given time. The desirability of the gifts was established by pilot tests. These gifts were also placed on the table. The experimenters sat behind the children, so the children would not be distracted by their presence and could concentrate on the task. All children were videotaped while working on the task.

Coding anger regulation strategies

To assess the children's anger regulation strategies, their behavior during the anger-eliciting task was coded based on the categorization of adaptive and maladaptive anger regulation strategies. The development and classification of these strategies as adaptive or maladaptive is explained in detail in Rohlf and Krahe (2015). At both time points, four maladaptive anger regulation strategies (*visual focus on the frustrating stimuli*, *verbal focus on the frustrating stimuli*, *venting the anger*, and *resignation*) and one adaptive strategy (*solution orientation*) were coded for each child. Each strategy was represented by at least one specific codable behavior (subcategory) that was assumed to reflect the superordinate strategy (see Table 1). For the majority of strategies, event-coding was used, counting the number of occurrences of each subcategory within the observation period (exceptions are described below). Coding was conducted by trained raters using the software Eudico Linguistic Annotator (ELAN; Wittenburg, Brugman, Russel, Klassmann, & Sloetjes, 2006).

The definition of the four maladaptive strategies was the same at both time points. *Visual focus on the frustrating stimuli* included eye movement toward potential frustrating objects, that is, the unreachable gifts and the running timer. *Verbal focus on the frustrating stimuli* included any verbal comments referring to potentially frustrating objects, that is, the running time, the unreachable gifts, the task itself, and the self (e.g., "I can't do it"). *Venting the anger* described any anger expressive behavior, including verbal expressions (e.g., swearing), facial and gestural expressions (e.g., clenching one's fist), and rough handling of the material (e.g., smashing the blocks or dice on the table). *Resignation* included refusing to continue working on the task for at least 3 seconds.

Due to some changes of the anger-eliciting task in material and procedure, the only adaptive strategy *solution orientation* was defined slightly differently at both time points: At T1, the sub-category "Testing an alternative strategy" was event-coded, counting the numbers of new attempts to complete the task. At T2, we counted the first occurrences of five predefined strategies, resulting in a scale ranging from 0 (*showed none of these strategies*) to 5 (*showed all of these strategies*). For the sub-category "Balancing out the tower", the time (in seconds) that the children spent trying to balance the tower was measured at T1. The dice tower at T2 was far less stable, so the tower collapsed much more often, and the duration of each balancing attempt could not be measured accurately. Instead, we counted the *number* of attempts to balance the dice tower (event-coded). At both time points, the sub-category "Working in a focused way" was rated by the coder on a scale from 1 (*very little engagement with the task*) to 4 (*very much engagement with the task*). Since at T2 the children did not have to copy the tower depicted on the picture, but were free to stack the dice as they wanted, they had more opportunities to show solution-oriented behavior. Thus, the number of times they rearranged the dice was event-coded as an additional category, and we also coded whether the children followed the example picture (0 vs. 1).

Sum scores of the corresponding sub-categories were calculated for the three maladaptive strategies *visual focus on frustrating stimuli*, *verbal focus on frustrating stimuli*, and *venting the anger*. A dichotomized score was created for *resignation* (0 vs. 1).² For the

²Originally, event-coding was used for resignation at both time points. At T2, the maximum number of resignations was one. To improve model fit, resignation T2 was declared as a categorical variable in all latent analyses. For the sake of consistency over time and to be able to test measurement invariance, we retrospectively changed the T1 scoring to this dichotomized score. The data of only three children at T1 were affected by this modification, with their score changing from 2 to 1.

Table 1. Coding scheme and inter-rater reliability.

Strategy	α_{T1}	α_{T2}	Sub-categories
Visual focus on frustrating stimuli	.71	.84	Looking at the timer Looking at the gifts
Verbal focus on frustrating stimuli	.92	.85	Talking negatively about the time Talking negatively about the gifts Talking negatively about the task Talking negatively about the self
Venting the anger	.73	.74	Verbal expression of anger Anger expression in facial expression and gesture Handling the material roughly
Resignation	.99	.92	Refusing to continue for at least 3 sec (0 vs. 1)
Solution orientation	.79	.89	T1 Testing an alternative strategy Balancing out the tower (in sec) Working in a focused way (1–4) T2 Testing an alternative strategy (0–5) Balancing out the tower Working in a focused way (1–4) Rearranging the dice Following the example picture (0 vs. 1)

Note. α = Krippendorff's alpha.

adaptive strategy *solution orientation*, the scaling of the subordinate categories differed, as explained above. Therefore, scores for each sub-category were first *z*-standardized and then averaged into an overall score.

A sub-sample of videos ($n = 121$ at T1 and $n = 120$ at T2) were double-coded by an independent rater to estimate the reliability of the coding process. The specific coding scheme as well as reliability information for both time points are presented in Table 1. We calculated Krippendorff's alpha as a measure of reliability of the coding. This coefficient is appropriate for coded data and determines the agreement between coders (Hayes & Krippendorff, 2007). Krippendorff's alphas ranged from .71 to .99 at T1 and from .74 to .92 at T2, respectively, indicating acceptable to good interrater agreement for both time points.

Emotional reactions to the anger-eliciting task

To check the success of our task in eliciting angry feelings in contrast to other negative emotions, children rated their experience of anger and sadness during the tower-building task on two items using a scale from 1 (*not at all angry/sad*) to 3 (*very angry/sad*). At T1, this assessment took place immediately after the tower-building task. At T2, it was placed after the assignment of the number of dice to the other child as a measure of aggressive behavior (explained in the following section).

Validation constructs

To establish criterion validity, we assessed several constructs at T1 and T2 that are assumed to be associated with maladaptive anger regulation. In particular,

we considered children's aggressive behavior based on three independent sources of information (teacher-report, child-report, and *in-situ* behavior), peer problems (teacher-, child-, and parent-report), and conduct problems (parent-report). These constructs (with the exception of *in-situ* behavior) were operationalized using Likert-type scales with two to five response options, that is they were measured at an ordinal level of measurement. Thus, we calculated Ordinal alpha instead of traditional Cronbach's alpha as an estimation of reliability (Gadermann, Guhn, & Zumbo, 2012).

Aggressive behavior: Teacher-report

At both data waves, teachers indicated the frequency of children's aggressive behavior in the last 6 months. We used six items adapted from the Children's Social Behavior Scale - Teacher Form (CSBS-T; Crick, 1996; e.g., "How often did this child hit, shove, or push peers"; $\alpha_{T1} = .94$, $\alpha_{T2} = .95$). The response scale ranged from 1 (*never*) to 5 (*daily*).

Aggressive behavior: Child-report

At T2, the children were presented the same six CSBS-items as the teachers in a reformulated, age-adapted version (e.g., "How often did you hit, shove, or push other children"; $\alpha = .74$). The response scale ranged from 1 (*never*) to 4 (*almost daily*).

Aggressive behavior in situ

At T2, a behavioral measure of aggression was developed that was derived from the dice-stacking task. Immediately after the task, the children were told that they could decide how many dice another child in another school would receive to complete the same

task. Participants were told that they could choose any number between 2 and 12 dice and were reminded that they had been given seven dice. The number of additional dice (>7) assigned to the alleged other child was taken as the measure of aggressive behavior. This operationalization of aggressive behavior is based on the same rationale as the Tangram Help/Hurt Task by Saleem, Anderson, and Barlett (2015). Because the task gets more difficult the more dice have to be used to build the tower, assigning a greater number of dice to another child than the participant had to handle him/herself can be viewed as reflecting an intention to harm, thus meeting the definition of aggressive behavior (Baron & Richardson, 1994). Accordingly, we transformed the raw number of dice into a scale using 7 dice as the reference point. All assignments of equal or less than 7 received a score of zero since assigning as many or fewer dice to the other child than they had received themselves represents a nonaggressive response. The resulting aggression scale ranged from 0 (7 dice or fewer) to 5 (maximum number of 12 dice).

Peer problems: Teacher-report

At T1 and T2, teachers rated each child on the degree of experienced peer problems, using three items ($\alpha_{t1} = .85$, $\alpha_{t2} = .87$). Two were taken from the Peer Problems subscale of the Strength and Difficulties Questionnaire (SDQ; Goodman, 1997; “is picked on or bullied by other children” and “is generally liked by other children,” reverse coding). The third item was self-constructed for taking into account the school context (“is often excluded when classmates play together at break time”). The response scale ranged from 1 (*not true*) to 3 (*certainly true*).

Peer problems: Child-report

At both time points, children were asked to report on their problems with peers in their class. At T1, peer problems were assessed using eight items ($\alpha = .80$). Of these, five items were taken from the Social Integration subscale of the Questionnaire on Social and Emotional Experiences at School of Elementary School Children (Fragebogen zur Erfassung emotionaler und sozialer Schulerfahrungen, FEES; Rauer & Schuck, 2003; e.g., “The other children often laugh at me”), and three items were taken from the Peer Acceptance subscale of the German version of the Harter-Scales (Asendorpf & van Aken, 1993; e.g., “I am liked by other children”, reverse coding). The response options at T1 were 1 (*no*) and 2 (*yes*). At T2, peer problems were assessed with seven items of the FEES (including the same as at T1; $\alpha = .87$). The response options at T2 were 1 (*no*), 2 (*rather no*), 3 (*rather yes*), and 4 (*yes*).

Peer problems and conduct problems: Parent-report

At both time points, parents completed the Peer Problems ($\alpha_{t1} = .79$, $\alpha_{t2} = .83$) and the Conduct Problems subscales ($\alpha_{t1} = .79$, $\alpha_{t2} = .79$; e.g., “often lies or cheats”) of the SDQ (Goodman, 1997). The response scale ranged from 1 (*not true*) to 3 (*certainly true*).

Data analysis plan

We used Mplus (Version 7.3, Muthén & Muthén, 1998–2015) for our analyses. Due to the categorical nature of most indicators, the WLSMV estimator was used in all latent analyses (Li, 2015). Potential age and gender differences in latent analyses were examined by applying Wald tests. Age groups were defined by splitting the sample in a younger and an older sub-sample by median ($Md = 8.02$ at T1).

Only 7.5% of the T1 sample did not participate at T2, and few differences were found between the participants who remained in the sample and those who dropped out. Dropouts were older ($t[597] = -2.18$, $p = .030$, $d = 0.34$, 95% CI [0.03, 0.65]), and were described by their teachers as more aggressive at T1 ($t[46.86] = -2.02$, $p = .049$, $d = 0.35$, 95% CI [0.04, 0.66]). In terms of missing responses, a total of 13% of data were missing, ranging from 0% to 34.2% across the examined variables. The highest missing rate of 34.2% was for teacher reports at T2 since not all teachers returned the questionnaires.

To deal with the missing data and to reduce biases due to selective dropout, multiple imputation was used (Asendorpf, van de Schoot, Denissen, & Hutteman, 2014; Rubin, 1987). This was done under the missing at random (MAR) assumption. We created 60 multiply imputed datasets in 120 iterations using R 3.2.1 (R Core Team, 2015), the default settings of the mice 2.25 package (van Buuren & Groothuis-Oudshoorn, 2011), and fully conditional specification (van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006). For the sake of consistency, we also imputed missing data at T1, resulting in slight deviations from analyses reported in previous studies using the T1 data (Kirsch et al., 2015; Rohlf, Busching, & Krahé, *in press*; Rohlf & Krahé, 2015). In addition to the variables used in the present study, we also included auxiliary variables (the remaining SDQ subscales) in the imputation model to improve parameter estimation (Yoo, 2009). All analyses reported in the Results section were based on the imputed data set.

Results

Manipulation check

The tower-building task was designed to induce angry feelings in the children in order to observe actual use of

anger regulation strategies. At both time points, anger ratings were above the midpoint of the response scale, and the tasks elicited significantly more anger than sadness (T1: $M_{\text{anger}} = 2.31$, $SD_{\text{anger}} = 0.64$ vs. $M_{\text{sadness}} = 1.84$, $SD_{\text{sadness}} = 0.72$; $t[598] = 11.97$, $p < .001$, $d = 0.94$, 95% CI [0.89, 1.00]; T2: $M_{\text{anger}} = 2.22$, $SD_{\text{anger}} = 0.59$ vs. $M_{\text{sadness}} = 1.65$, $SD_{\text{sadness}} = 0.61$; $t[598] = 15.43$, $p < .001$, $d = 1.24$, 95% CI [1.19, 1.29]). This indicates that the anger induction was successful. In addition, experienced anger was not correlated with participant's age ($r_s \leq |-.02|$, $p_s \geq .643$). This indicates that our task was able to elicit anger regardless of age. Regarding gender, boys reported more anger than did girls at T1 ($M_{\text{girls}} = 2.25$, $SD_{\text{girls}} = 0.63$ vs. $M_{\text{boys}} = 2.38$, $SD_{\text{boys}} = 0.64$; $t[597] = 2.46$, $p = .014$, $d = 0.20$, 95% CI [0.04, 0.36]), but there was no difference at T2 ($t[597] = 0.97$, $p = .330$, $d = 0.08$, 95% CI [-0.08, 0.24]).

As a manipulation check for the *in-situ* measure of aggressive behavior used at T2 (number of dice), a sub-sample of $n = 76$ randomly selected children were asked whether they thought the task was more difficult when more dice had to be used. The vast majority of these children (86%) answered in the affirmative. This result supports the assumption that participants were aware of the harmful nature of assigning more dice to the alleged other child, suggesting that assigning a higher number of dice than they had received themselves can be interpreted as a measure of aggressive behavior, defined by the intention to harm (Krahé, 2013).

Descriptive statistics, gender and age differences, factor analysis, and correlations

Descriptive statistics of the use of anger regulation strategies at T1 and T2 as well as tests for gender differences are presented in Table 2. *Visual focus* and *venting the anger* were the most frequently used strategies, *resignation* had the lowest frequency. Few gender differences emerged in the use of anger regulation strategies: Boys used more *visual focus* than did girls at T1 ($t[597] = 3.19$, $p = .002$, $d = 0.26$, 95% CI [0.10, 0.42]), and more *venting the anger* at T2 ($t[597] = 2.57$, $p = .010$, $d = 0.21$, 95% CI [0.05, 0.37]).

Table 3 presents stability information, factor loadings, bivariate correlations of the anger regulation strategies at T1 and T2, and their correlations with age. Stabilities over the course of the 3 years between T1 and T2 were low to moderate, r_s ranging from .12 to .29. *Solution orientation* showed the lowest stability due to the redefinition of that indicator, as explained above. Pearson's correlation coefficients among the strategies were low to moderate at both time points. Age at T1 was negatively correlated with *visual* and *verbal focus*,

and positively with *solution orientation* at both time points, suggesting that maladaptive anger regulation decreases with age. The measurement model for maladaptive anger regulation at T2 showed a good fit after freeing residual covariances between the manifest indicators *venting the anger* and *verbal focus*, *resignation* and *solution orientation*, and *resignation* and *visual focus* ($\chi^2[2] = 3.88$, $p = .144$; RMSEA = .03; WRMR = 0.25; CFI = .99; TLI = .97). The measurement model for maladaptive anger regulation T1 also showed a good fit after freeing residual covariances between *visual focus* and *solution orientation*, and *resignation* and *solution orientation* ($\chi^2[3] = 11.84$, $p = .008$; RMSEA = .07; WRMR = 0.59; CFI = .99; TLI = .96). The factor loadings indicate that the maladaptive regulation strategies were positively associated with each other, whereas the single adaptive strategy *solution orientation* was negatively associated with the other strategies. This confirms the theoretical classification of the strategies as adaptive and maladaptive.

Descriptive statistics, gender differences, stabilities and correlations of the validation constructs with age are presented in Appendix C. Stability of the constructs were low to moderate across the time interval of 3 years (r_s ranging from .29 to .64, $p_s < .001$). All constructs were positively correlated with each other. The significant cross-sectional correlations of our *in-situ* measure of aggressive behavior (number of dice) with child- ($r = .22$, $p < .001$, 95% CI [.14, .30]) and teacher-reported ($r = .15$, $p < .001$, 95% CI [.07, .23]) aggressive behavior provide further support for its validity as a measure of aggressive behavior.

Measurement invariance

To demonstrate the equivalence of the two tasks measuring anger regulation at T1 and T2, longitudinal measurement invariance was tested (Millsap & Cham, 2012). Table 4 presents a summary of the corresponding analysis. A change of $<-.010$ in CFI, and $<.015$ in RMSEA was interpreted as supporting the invariance assumption (Chen, 2007). For establishing configural and higher levels of invariance, the residual covariance of the strategy *visual focus* at T1 and T2 was freed to improve model fit. Full invariance across all indicators could only be established at the configural level, but not for the higher levels. Steenkamp and Baumgartner (1998) recommend testing partial measurement invariance when full invariance cannot be applied. Thus, for further invariance tests, the strategy *solution orientation* was freed between the two time points, since this strategy had changed in definition over time. In the end, the highest level of invariance for our proposed

Table 2. Descriptive statistics and gender differences of anger regulation strategies.

	T1					T2				
	Range	Total M (SD)	Girls M (SD)	Boys M (SD)	Gender difference	Range	Total M (SD)	Girls M (SD)	Boys M (SD)	Gender difference
Visual focus	0–39	4.07 (3.67)	3.60 (3.12)	4.55 (4.11)	$t = 3.19^{**}$ $d = 0.26$	0–14	3.45 (2.98)	3.38 (2.93)	3.52 (3.02)	$t = 0.58$ $d = 0.05$
Verbal focus	0–27	2.75 (3.54)	2.62 (3.57)	2.88 (3.50)	$t = 0.90$ $d = 0.07$	0–10	0.81 (1.63)	0.74 (1.56)	0.90 (1.69)	$t = 1.20$ $d = 0.10$
Venting	0–22	4.33 (3.87)	4.36 (3.90)	4.29 (3.84)	$t = -0.22$ $d = 0.02$	0–14	2.29 (2.73)	2.01 (2.54)	2.58 (2.88)	$t = 2.57^*$ $d = 0.21$
Resignation	0–2	0.02 (0.14)	0.02 (0.15)	0.02 (0.14)	$t = 0.25$ $d = 0.02$	0–1	0.07 (0.26)	0.07 (0.25)	0.07 (0.26)	$t = 0.29$ $d = 0.02$
Solution orientation ^a	–	0.00 (1.60)	-0.12 (1.51)	0.13 (1.68)	$t = 1.92$ $d = 0.16$	–	-0.04 (0.51)	0.00 (0.50)	-0.08 (0.51)	$t = -1.94$ $d = 0.16$

Note. ^aScores for solution orientation were z-transformed; $n_{girls} = 304$, $n_{boys} = 295$; df for all independent t -tests was 597.
* $p < .05$. ** $p < .01$.

Table 3. Stabilities, correlations with age, factor loadings, and intercorrelations of anger regulation strategies.

	Visual focus T1	Verbal focus T1	Venting T1	Resignation T1	Solution T1	Visual focus T2	Verbal focus T2	Venting T2	Resignation T2	Solution T2
Age T1	-.19***	-.21***	-.13**	-.06	.34***	-.14***	-.13**	-.02	.02	.10*
Visual focus T1		.34***	.11**	.08*	-.35***	.25***	.11**	.04	.09*	-.09*
Verbal focus T1			.43***	.16***	-.43***	.07	.29***	.23***	.02	-.05
Venting T1				.14***	-.27***	.05	.21***	.22***	-.01	-.02
Resignation T1					-.32***	.05	.15***	.12**	.17***	-.14***
Solution T1						-.09*	-.20***	-.13**	-.05	.12**
Visual focus T2							.19***	.12**	.30***	-.21***
Verbal focus T2								.44***	.16***	-.16***
Venting T2									.15***	-.13**
Resignation T2										-.40***
Solution T2										
Factor loading	.36***	.91***	.47***	.38***	-.49***	.47***	.37***	.30***	.62***	-.44***

Notes. Solution = solution orientation; stabilities of strategies are highlighted in bold.
* $p < .05$. ** $p < .01$. *** $p < .001$.

behavioral observation method was partial metric invariance, as indicated by changes in fit indices below the critical threshold ($\Delta CFI = -.004$; $\Delta RMSEA = -.001$). That is, the factor loadings of all indicators—except for the strategy *solution orientation*—could be constrained to be equal across a time interval of 3 years. Additionally constraining the intercepts of the indicators, a condition for scalar invariance, worsened the model fit, therefore partial scalar invariance could not be assumed.

In addition to measurement invariance over time, we tested measurement invariance across gender groups to examine the equivalence of the method for boys and girls (Widaman & Reise, 1997; see Table 5). At both T1 and T2 full scalar invariance could be established

($\Delta CFI_{T1} = -.004$, $\Delta RMSEA_{T1} = -.007$; $\Delta CFI_{T2} = -.007$, $\Delta RMSEA_{T2} = .001$). That is, the factor loadings and the intercepts of all indicators could be constrained to be equal across boys and girls. Regarding the role of age, we tested the measurement invariance across the younger and older sub-sample of our study (see Table 5). At T1, partial metric invariance could be established (the indicator verbal focus was freed; $\Delta CFI = .000$, $\Delta RMSEA = -.014$). At T2, full scalar invariance could be established ($\Delta CFI = -.009$, $\Delta RMSEA = .012$).

Stability

To assess construct stability, a latent-state model was computed to test the correlation between the two latent

Table 4. Measurement invariance of maladaptive anger regulation over time.

Level of invariance	χ^2	df	TLI	WRMR	CFI	ΔCFI	RMSEA	$\Delta RMSEA$
Configural	81.75	27	.928	0.999	.957	–	.058	–
Full metric	139.62	31	.876	1.447	.915	-.042	.076	.018
Full scalar	213.76	35	.820	1.802	.860	-.055	.092	.016
Full strict	381.18	39	.690	2.486	.732	-.128	.121	.029
Partial metric	89.55	30	.930	1.111	.953	-.004	.057	-.001
Partial scalar	137.34	33	.888	1.407	.918	-.035	.073	.016
Partial strict	190.41	36	.848	1.707	.879	-.039	.085	.012

Note. All χ^2 -statistics are significant at $p < .001$.

Table 5. Measurement invariance of maladaptive anger regulation across gender and age.

Level of invariance	χ^2	<i>df</i>	<i>p</i>	TLI	WRMR	CFI	Δ CFI	RMSEA	Δ RMSEA
Gender T1									
Configural	18.17	6	.006	.949	0.744	.985	–	.081	–
Full Metric	29.63	11	.002	.957	1.275	.977	–.008	.074	–.007
Full Scalar	37.69	16	.002	.966	1.459	.973	–.004	.067	–.007
Full Strict	84.11	20	.000	.919	2.320	.919	–.054	.103	.036
Gender T2									
Configural	4.94	4	.294	.986	0.302	.995	–	.027	–
Full Metric	13.62	9	.137	.971	0.804	.986	–.009	.036	.009
Full Scalar	21.30	14	.094	.970	1.035	.979	–.007	.037	.001
Full Strict	30.22	18	.035	.960	1.270	.964	–.015	.043	.006
Age T1									
Configural	12.462	6	.052	.976	0.599	.990	–	.059	–
Full Metric	44.60	11	.000	.906	1.582	.948	–.042	.101	.042
Full Scalar	121.03	16	.000	.798	2.704	.838	–.110	.148	.047
Full Strict	259.92	20	.000	.630	4.162	.630	–.208	.200	.052
Partial Metric	16.63	10	.092	.981	0.844	.990	.000	.045	–.014
Partial Scalar	95.65	14	.000	.820	2.224	.874	–.116	.139	.094
Partial Strict	212.31	17	.000	.645	3.479	.698	–.176	.196	.057
Age T2									
Configural	4.70	4	.320	.990	0.283	.996	–	.024	–
Full Metric	10.00	9	.351	.995	0.648	.994	–.002	.018	–.006
Full Scalar	19.37	14	.151	.980	0.975	.985	–.009	.030	.012
Full Strict	46.33	18	.000	.916	1.628	.924	–.061	.071	.041

Note. Partial invariance for age T1 established by freeing the indicator verbal focus.

factors of maladaptive anger regulation at T1 and T2. This model was run under the assumption of partial metric longitudinal invariance as explained above (for model fit information, see Table 4). The latent correlation was $r = .52$ ($p < .001$, 95% CI [.37, .67]). This result indicates moderate stability of the latent factor of maladaptive anger regulation over the course of 3 years. The stability did not differ between girls and boys ($W = 0.03$, $p = .873$), but between the age groups ($W = 7.70$, $p = .006$). Younger children showed higher construct stability than did older children ($r_{\text{younger}} = .59$ vs. $r_{\text{older}} = .40$; all $ps < .001$).

Criterion validity

Concurrent criterion validity of T1 and T2 maladaptive anger regulation was assessed by cross-sectional correlations with the criterion measures at both time points. Predictive criterion validity for behavioral observation

at T1 was assessed by correlating it with the T2 criterion measures. All correlations are partial correlations, controlled for age and gender. All criterion constructs (with the exception of the *in-situ* measure of aggressive behavior, which was a single-item measure) were modelled as latent variables. The resulting structural equation models testing concurrent and predictive validity showed good to adequate model fit, with $.912 < CFI < .997$ and $.035 < RMSEA < .073$ (van de Schoot, Lugtig, & Hox, 2012). Table 6 presents the results for concurrent and predictive validity. As expected, maladaptive anger regulation showed positive correlations with aggressive behavior, peer problems, and conduct problems, both cross-sectionally at both time points and prospectively from T1 to T2. The only exception was the nonsignificant prospective association of T1 anger regulation with T2 teacher-rated aggressive behavior, which needs to be interpreted in view of the fact that children were rated by different teachers at

Table 6. Concurrent and predictive validity of the behavioral observation measures.

	Maladaptive anger regulation					
	Concurrent T1		Concurrent T2		Predictive T1 to T2	
	<i>r</i>	95% CI	<i>r</i>	95% CI	<i>r</i>	95% CI
Aggressive behavior						
Teacher-report	.20***	[.08, .31]	.22*	[.02, .42]	.06	[–.06, .17]
Child-report	– ^a	–	.40***	[.17, .63]	.23***	[.10, .36]
<i>In-situ</i> behavior	– ^a	–	.22**	[.07, .36]	.11*	[.01, .20]
Peer problems						
Teacher-report	.17*	[.04, .30]	.42***	[.23, .62]	.27***	[.15, .40]
Child-report	.27***	[.15, .38]	.40***	[.23, .58]	.15*	[.03, .26]
Parent-report	.19**	[.06, .31]	.32**	[.12, .52]	.21**	[.08, .33]
Conduct problems						
Parent-report	.21***	[.09, .33]	.40***	[.19, .61]	.27***	[.14, .40]

Notes. ^aNot assessed at T1; all correlations are controlled for age and gender.

* $p < .05$. ** $p < .01$. *** $p < .001$.

the two data waves. The strongest associations, with correlations mostly around $r = .40$, were found for the concurrent associations of anger regulation with the validation constructs at T2. Wald tests revealed no gender differences in these correlations ($W_s \leq 2.10$, $ps \geq .147$), with the exception of the predictive validity of *in-situ* aggressive behavior ($W = 4.57$, $p = .033$). The prospective correlation between maladaptive anger regulation T1 and *in-situ* aggression T2 was only significant in girls ($r = .23$, $p < .001$), but not in boys ($r = -.01$, $p = .871$). No significant age group differences were found ($W_s \leq 3.31$, $ps \geq .069$). In sum, the findings provide conclusive support for the validity of the behavioral observation measures.

Discussion

The aim of this study was to further develop the behavioral observation assessment of anger regulation proposed by Rohlf and Krahe (2015) for the use in longitudinal research on the development of anger regulation in childhood and its role as a risk factor for aggressive behavior, peer problems, and conduct problems. To elicit anger, children in the age group of 6 to 10 years were given a practically unsolvable dexterity task at T1 (building a tower using wooden blocks). About 3 years later at T2, the same children were asked to work on a slightly different, but age-appropriate variation of the task that was supposed to be conceptually equivalent (building a tower by stacking dice). Both tower-building tasks were successful in eliciting angry feelings, which made them suitable for the observational study of anger regulation *in situ*.

The same coding scheme was used at both data waves to categorize specific anger regulation strategies. The reliability of our new task was good, as indicated by high interrater agreement. Only few gender differences in the use of regulation strategies were found, and these were in line with previous studies (e.g., boys vented their anger more frequently than did girls at T2; for a review, see Kerr & Schneider, 2008). Tests of measurement invariance revealed partial metric invariance as the highest level of longitudinal invariance. Metric invariance indicates that the underlying latent factor has the same unit of measurement at both time points, which is a prerequisite for the comparison of regression slopes in longitudinal research (Chen, 2007). Furthermore, tests of measurement invariance across gender and age groups indicated that the factorial structure of both observation measures were similar for girls and boys, and younger and older children, respectively.

Regarding stability, the latent construct of anger regulation showed a significant and substantial correlation

between the two time points. This result indicates moderate construct stability over time, and provides further evidence of the equivalence of the two measures. Moreover, the substantial stability is consistent with the conceptualization of emotion regulation as a stable individual difference variable (Cole et al., 1994).

To establish the validity of our behavioral observation measure, we assessed problematic interpersonal outcomes associated with maladaptive emotion regulation in general (criterion validity). Specifically, we examined the correlations of maladaptive anger regulation cross-sectionally (concurrent validity) as well as longitudinally (predictive validity) with aggressive behavior, peer problems, and conduct problems. These validation constructs were assessed by a multi-method approach drawing on self-, teacher-, and parent-reports as well as *in-situ* behavior. For both concurrent and predictive analyses, we were able to demonstrate positive and for the most part significant correlations between maladaptive anger regulation and the validation constructs. In addition, the correlations were similar in girls and boys, as well as in younger and older children in our sample. These findings support the validity of our measure as they are in line with a large body of research also showing that deficits in anger regulation are related to behavioral and peer problems (e.g., Dearing et al., 2002).

Overall, our study was successful in developing an equivalent of the original behavioral observation assessment of Rohlf and Krahe (2015) that takes developmental changes in the children's cognitive and motor skills as well as interests into account. Such equivalent measurement tools are required in the longitudinal study of anger regulation in childhood, covering substantial lengths of time.

Strengths and limitations

We believe that our study has several strengths. It is based on the behavioral observation of anger regulation in a large sample of children who were assessed twice over a 3-year period. Our two assessments were found to be conceptually equivalent and both measures worked equally well for boys and girls. A range of outcome measures was included to establish criterion validity of the measures at T1 and T2, using data from multiple informants. Moreover, the anger-eliciting task used at T2 had the advantage of yielding a behavioral measure of aggression *in situ*, the number of dice assigned to another child. This feature facilitates a direct mapping of anger regulation strategies onto aggressive responses within the same situational context.

At the same time, some limitations have to be acknowledged. First, since we used behavioral

observation, our conclusions and inferences can be applied only to anger regulation strategies that can be inferred from overt behavior. Like any behavioral observation, our measure cannot capture mental strategies, such as cognitive reappraisal, that are not necessarily reflected in overt behavior. As children get older, cognitive strategies may become more and more important in emotion regulation processes (e.g., McRae et al., 2012). Therefore, for future studies we suggest a combination of behavioral observation with self-report measures to obtain a clearer picture of children's regulation strategies.

Second, the children were exposed to an arranged anger-eliciting situation that constrained the possible regulation strategies they could show. Specifically, children did not have the opportunity to select or modify the situation, two important regulation approaches (Gross, 1998, 2014). Another effective and frequently used adaptive strategy in natural situations is distraction from the frustrating stimuli (e.g., Denson, Moulds, & Grisham, 2012). However, in our tasks, the children had very limited opportunities to withdraw from the task and distract themselves.

Another limitation is that we only assessed criterion validity (concurrent and predictive). However, for the T1 task, construct validity could be demonstrated by relating it to the conceptually related construct of anger reactivity (Rohlf & Krahé, 2015). Given the construct validity of our T1 measure and the conceptualization of emotion regulation as a stable individual difference variable (Cole et al., 1994), the substantial temporal stability from T1 to T2 may be interpreted as an indication that our T2 task also has construct validity as a measure of anger regulation (Cronbach & Meehl, 1955). As a final limitation, we only assessed the experience of two emotional states, namely anger and sadness, during the tower-building task. Other affective states such as anxiety or joy should be considered in future studies.

Despite these limitations, the behavioral observation measure offers an age-appropriate, economic, and valid approach for assessing the use of anger regulation strategies in an actual anger-inducing situation in an equivalent way at different points of development in childhood. With carefully trained experimenters, the anger-eliciting tasks are easy to administer and can be completed in a short time. The categories are well-defined and grounded in theory, capture a range of different strategies, and yield reliable codings, as indicated by the high inter-coder reliability. They add a methodological tool to the study of anger regulation that lends itself to the longitudinal analysis of both adaptive and maladaptive forms of anger regulation in middle childhood. At a more general level, they demonstrate an approach for

designing conceptually and empirically equivalent measures to study developmental processes over time.

Acknowledgments

The authors would like to thank Ronja Fink and Eva Bausch for their assistance with the coding process.

ORCID

Fabian Kirsch  <http://orcid.org/0000-0003-3894-5635>

References

- Aldao, A., Nolen-Hoeksema, S., & Schweizer, S. (2010). Emotion-regulation strategies across psychopathology: A meta-analytic review. *Clinical Psychology Review, 30*, 217–237. doi:10.1016/j.cpr.2009.11.004
- Asendorpf, J. B., & van Aken, M. A. G. (1993). Deutsche Versionen der Selbstkonzeptskalen von Harter [German version of the Harter's self-concept scales]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 25*, 64–86.
- Asendorpf, J. B., van de Schoot, R., Denissen, J. J. A., & Hutteman, R. (2014). Reducing bias due to systematic attrition in longitudinal studies: The benefits of multiple imputation. *International Journal of Behavioral Development, 38*, 453–460. doi:10.1177/0165025414542713
- Baron, R. A., & Richardson, D. S. (1994). *Human aggression* (2nd ed.). New York, NY: Plenum Press.
- Bartolotta, T. E., & Shulman, B. B. (2010). Child development. In B. B. Shulman & N. C. Singleton (Eds.), *Language development. Foundations, processes, and clinical applications* (pp. 35–53). Sudbury, MA: Jones and Bartlett.
- Berkowitz, L. (2012). A different view of anger: The cognitive-neoassociation conception of the relation of anger to aggression. *Aggressive Behavior, 38*, 322–333. doi:10.1002/ab.21432
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*, 464–504. doi:10.1080/10705510701301834
- Cole, P. M., Michel, M. K., & Teti, L. O. D. (1994). The development of emotion regulation and dysregulation: A clinical perspective. *Monographs of the Society for Research in Child Development, 59*, 73–100. doi:10.2307/1166139
- Crick, N. R. (1996). The role of overt aggression, relational aggression, and prosocial behavior in the prediction of children's future social adjustment. *Child Development, 67*, 2317–2327. doi:10.2307/1131625
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302. doi:10.1037/h0040957
- Dearing, K. F., Hubbard, J. A., Ramsden, S. R., Parker, E. H., Relyea, N., Smithmyer, C. M., & Flanagan, K. D. (2002). Children's self-reports about anger regulation: Direct and indirect links to social preference and aggression. *Merrill-Palmer Quarterly, 48*, 308–336. doi:10.1353/mpq.2002.0011
- Denson, T. F., Moulds, M. L., & Grisham, J. R. (2012). The effects of analytical rumination, reappraisal, and distraction on anger experience. *Behavior Therapy, 43*, 355–364. doi:10.1016/j.beth.2011.08.001

- Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, *17*, 124–129. doi:10.1037/h0030377
- Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research & Evaluation*, *17*(3), 1–13.
- Goodman, R. (1997). The strengths and difficulties questionnaire: A research note. *Journal of Child Psychology and Psychiatry*, *38*, 581–586. doi:10.1111/j.1469-7610.1997.tb01545.x
- Gross, J. J. (1998). The emerging field of emotion regulation: An integrative review. *Review of General Psychology*, *2*, 271–299. doi:10.1037/1089-2680.2.3.271
- Gross, J. J. (2014). Emotion regulation: Conceptual and empirical foundations. In J. J. Gross (Ed.), *Handbook of emotion regulation* (2nd ed., pp. 3–20). New York, NY: Guilford Press.
- Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*, 77–89. doi:10.1080/19312450709336664
- Helmsen, J., & Petermann, F. (2010). Emotionsregulationsstrategien und aggressives Verhalten im Kindergartenalter [Emotion regulation strategies and aggressive behavior of preschool children]. *Praxis der Kinderpsychologie und Kinderpsychiatrie*, *59*, 775–791. doi:10.13109/prkk.2010.59.10.775
- Kerr, M. A., & Schneider, B. H. (2008). Anger expression in children and adolescents: A review of the empirical literature. *Clinical Psychology Review*, *28*, 559–577. doi:10.1016/j.cpr.2007.08.001
- Kirsch, F., Rohlf, H., & Krahé, B. (2015). Measuring anger regulation in middle childhood through behavioural observation: A longitudinal validation. *European Journal of Developmental Psychology*, *12*, 718–727. doi:10.1080/17405629.2015.1101375
- Krahé, B. (2013). *The social psychology of aggression* (2nd ed.). Hove, UK: Psychology Press.
- Li, C.-H. (2015). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*, 936–949. doi:10.3758/s13428-015-0619-7
- McRae, K., Gross, J. J., Weber, J., Robertson, E. R., Sokol-Hessner, P., Ray, R. D., ... Ochsner, K. N. (2012). The development of emotion regulation: An fMRI study of cognitive reappraisal in children, adolescents and young adults. *Social Cognitive and Affective Neuroscience*, *7*, 11–22. doi:10.1093/scan/nsr093
- Millsap, R. E., & Cham, H. (2012). Investigating factorial invariance in longitudinal data. In B. P. Laursen, T. D. Little, & N. A. Card (Eds.), *Handbook of developmental research methods* (pp. 109–127). New York, NY: Guilford Press.
- Morris, A. S., Silk, J. S., Steinberg, L., Terranova, A. M., & Kithakye, M. (2010). Concurrent and longitudinal links between children's externalizing behavior in school and observed anger regulation in the mother-child dyad. *Journal of Psychopathology and Behavioral Assessment*, *32*, 48–56. doi:10.1007/s10862-009-9166-9
- Muthén, L. K., & Muthén, B. O. (1998-2015). *Mplus user's guide: Statistical analysis with latent variables* (7th ed.). Los Angeles, CA: Muthén & Muthén.
- Parker, E. H., Hubbard, J. A., Ramsden, S. R., Relyea, N., Dearing, K. F., Smithmyer, C. M., & Schimmel, K. D. (2001). Children's use and knowledge of display rules for anger following hypothetical vignettes versus following live peer interaction. *Social Development*, *10*, 528–557. doi:10.1111/1467-9507.00179
- Rauer, W., & Schuck, K.-D. (2003). *FEES 3–4: Fragebogen zur Erfassung emotionaler und sozialer Schulerfahrungen von Grundschulkindern dritter und vierter Klassen [FEES 3–4: Questionnaire for the assessment of social and emotional school experiences of elementary school children in third and fourth grade]*. Göttingen, Germany: Beltz Test GmbH.
- R Core team. (2015). *R: A language and environment for statistical computing*. Vienna, Australia: Foundation for Statistical Computing.
- Rohlf, H., Busching, R., & Krahé, B. (in press). Longitudinal links between maladaptive anger regulation, peer problems, and aggression in middle childhood. *Merrill-Palmer Quarterly*.
- Rohlf, H., & Krahé, B. (2015). Assessing anger regulation in middle childhood: Development and validation of a behavioral observation measure. *Frontiers in Psychology*, *6*, 453. doi:10.3389/fpsyg.2015.00453
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley.
- Saleem, M., Anderson, C. A., & Barlett, C. P. (2015). Assessing helping and hurting behaviors through the Tangram Help/Hurt Task. *Personality and Social Psychology Bulletin*, *41*, 1345–1362. doi:10.1177/0146167215594348
- Schaie, K. W., & Hofer, S. M. (2001). Longitudinal studies in aging research. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (5th ed., pp. 53–77). San Diego, CA: Academic Press.
- Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*, 78–107. doi:10.1086/209528
- Trampe, D., Quoidbach, J., & Taquet, M. (2015). Emotions in everyday life. *PLoS ONE*, *10*, e0145450. doi:10.1371/journal.pone.0145450
- Underwood, M. K. (1997). Top ten pressing questions about the development of emotion regulation. *Motivation and Emotion*, *21*, 127–146. doi:10.1023/A:1024482516226
- Underwood, M. K., Coie, J. D., & Herbman, C. R. (1992). Display rules for anger and aggression in school-age children. *Child Development*, *63*, 366–380. doi:10.2307/1131485
- van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C. G., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, *76*, 1049–1064. doi:10.1080/10629360600810434
- van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*, 1–67. doi:10.18637/jss.v045.i03
- van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*, 486–492. doi:10.1080/17405629.2012.686740
- von Salisch, M. (2000). *Wenn Kinder sich ärgern: Emotionsregulation in der Entwicklung. [When children feel angry: Emotion regulation in development]*. Göttingen, Germany: Hogrefe.

- von Salisch, M., Zeman, J., Luepschen, N., & Kanevski, R. (2014). Prospective relations between adolescents' social-emotional competencies and their friendships. *Social Development, 23*, 684–701. doi:10.1111/sode.12064
- Waters, S. F., & Thompson, R. A. (2014). Children's perceptions of the effectiveness of strategies for regulating anger and sadness. *International Journal of Behavioral Development, 38*, 174–181. doi:10.1177/0165025413515410
- Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention. Methodological advances from alcohol and substance abuse research* (pp. 281–324). Washington, DC: American Psychological Association.
- Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). *ELAN: A professional framework for multimodality research*. Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation, Genoa.
- Wittmann, M., Arce, E., & Santisteban, C. (2008). How impulsiveness, trait anger, and extracurricular activities might affect aggression in school children. *Personality and Individual Differences, 45*, 618–623. doi:10.1016/j.paid.2008.07.001
- Wohlwill, J. F. (1973). *The study of behavioral development*. New York, NY: Academic Press.
- Yoo, J. E. (2009). The effect of auxiliary variables and multiple imputation on parameter estimation in confirmatory factor analysis. *Educational and Psychological Measurement, 69*, 929–947. doi:10.1177/0013164409332225
- Zeman, J., & Shipman, K. (1997). Social-contextual influences on expectancies for managing anger and sadness: The transition from middle childhood to adolescence. *Developmental Psychology, 33*, 917–924. doi:10.1037/0012-1649.33.6.917

Appendix A

The participants were part of a larger sample of 1,657 children from 33 public elementary schools who participated in a longitudinal study on intrapersonal developmental risk factors in childhood and

adolescence. Children were videotaped if their parents gave permission ($n = 1,183$). It was not possible to code all videos due to limited resources. Therefore, a subsample of 599 children was randomly chosen for the T1 coding procedure. The resulting subsample did not differ significantly on any of the T1 variables included in the present study (see Table A1).

Appendix B

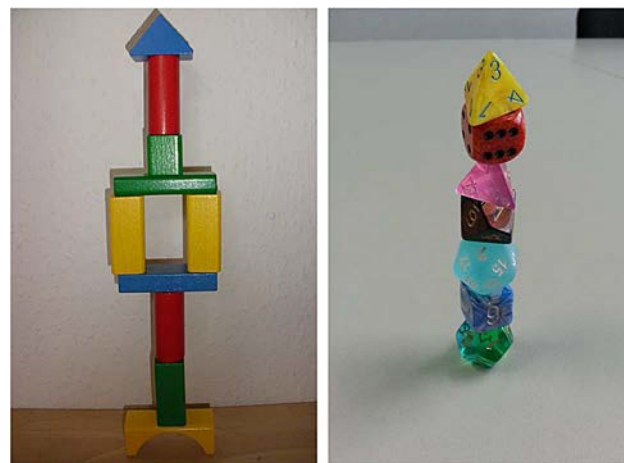


Figure B1. Photos for the tower-building task used at T1 (left) and T2 (right).

Table A1. Differences between included and not included children on T1 variables.

	Range	Included		Not included		Difference
		<i>N</i>	<i>M</i> (<i>SD</i>)	<i>N</i>	<i>M</i> (<i>SD</i>)	
Aggressive behavior						
Teacher-report	1–5	585	1.51 (0.68)	806	1.48 (0.66)	$t(1389) = -0.84$ $d = 0.04$
Peer problems						
Teacher-report	1–3	536	1.22 (0.34)	747	1.25 (0.37)	$t(1281) = 1.10$ $d = 0.08$
Child-report	1–2	598	1.18 (0.19)	1044	1.18 (0.21)	$t(1640) = 0.22$ $d = 0.01$
Parent-report	1–3	554	1.22 (0.30)	765	1.22 (0.31)	$t(1317) = 0.37$ $d = 0.02$
Conduct problems						
Parent-report	1–3	557	1.31 (0.32)	762	1.29 (0.30)	$t(1317) = -1.13$ $d = 0.06$

Note. All differences between the two groups are nonsignificant.

Appendix C

Table A2. Descriptive statistics and gender differences of validation constructs.

	Range	T1				T2			
		Total M (SD)	Girls M (SD)	Boys M (SD)	Gender difference	Total M (SD)	Girls M (SD)	Boys M (SD)	Gender difference
Aggressive behavior									
Teacher-report	1–5	1.52 (0.68)	1.38 (0.58)	1.67 (0.75)	$t = 5.30^{***}$ $d = 0.43$	1.57 (0.78)	1.49 (0.70)	1.65 (0.84)	$t = 2.54^*$ $d = 0.21$
Child-report ^a	1–4	–	–	–	–	1.29 (0.35)	1.22 (0.29)	1.36 (0.39)	$t = 5.00^{***}$ $d = 0.41$
<i>In-situ</i> behavior ^a	0–5	–	–	–	–	0.73 (1.39)	0.74 (1.38)	0.72 (1.39)	$t = -0.18$ $d = 0.01$
Peer problems									
Teacher-report	1–3	1.22 (0.34)	1.20 (0.33)	1.24 (0.35)	$t = 1.44$ $d = 0.12$	1.36 (0.50)	1.35 (0.51)	1.36 (0.48)	$t = 0.25$ $d = 0.02$
Child-report	T1: 1–2 T2: 1–4	1.18 (0.19)	1.16 (0.19)	1.20 (0.20)	$t = 2.51^*$ $d = 0.21$	2.17 (0.33)	2.17 (0.33)	2.16 (0.33)	$t = -0.37$ $d = 0.03$
Parent-report	1–3	1.26 (0.35)	1.25 (0.34)	1.27 (0.36)	$t = 0.70$ $d = 0.06$	1.36 (0.44)	1.36 (0.44)	1.37 (0.45)	$t = 0.28$ $d = 0.02$
Conduct problems									
Parent-report	1–3	1.35 (0.37)	1.33 (0.37)	1.38 (0.37)	$t = 1.65$ $d = 0.14$	1.38 (0.39)	1.37 (0.40)	1.38 (0.38)	$t = 0.31$ $d = 0.03$

Notes. ^aNot assessed at T1; $n_{\text{girls}} = 304$, $n_{\text{boys}} = 295$; df for all independent t -tests was 597.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Table A3. Stabilities, correlations with age, and intercorrelations of the validation constructs.

	Aggr-T T1	Peer-T T1	Peer-C T1	Peer-P T1	Cond-P T1	Aggr-T T2	Aggr-C T2	Aggr-S T2	Peer-T T2	Peer-C T2	Peer-P T2	Cond-P T2
Age T1	.00	.12**	.06	.01	.02	.03	.13**	-.09*	.11**	.02	.05	.04
Aggr-T T1		.38***	.19***	.20***	.33***	.45***	.29***	.07	.33***	.23***	.28***	.32***
Peer-T T1			.26***	.31***	.26***	.22***	.21***	.05	.31***	.26***	.35***	.30***
Peer-C T1				.29***	.21***	.17***	.12**	.03	.27***	.29***	.32***	.25***
Peer-P T1					.52***	.31***	.21***	.15***	.35***	.29***	.64***	.43***
Cond-P T1						.35***	.22***	.15***	.38***	.26***	.43***	.61***
Aggr-T T2							.37***	.15***	.50***	.19***	.40***	.44***
Aggr-C T2								.22***	.25***	.18***	.35***	.30***
Aggr-S T2									.09*	.11**	.14***	.15***
Peer-T T2										.45***	.58***	.48***
Peer-C T2											.46***	.34***
Peer-P T2												.58***
Cond-P T2												

Notes. Aggr = aggressive behavior; Peer = peer problems; Cond = conduct Problems; T = teacher-report; C = child-report; P = parent-report; stabilities of constructs are presented in bold. * $p < .05$. ** $p < .01$. *** $p < .001$.