Child & Youth Care Forum
https://doi.org/10.1007/s10566-019-09517-5

ORIGINAL PAPER

Development and Preliminary Evaluation of the Reaction to Unacceptable Behavior Inventory: A Questionnaire to Measure Progress in Implementation of Non-violent Resistance

K. M. Visser[1] · K. van Gink[1,2] · F. Thissen[3] · T. A. Visser[4] · T. Rimehaug[5,6] · L. C. M. Jansen[1] · A. Popma[1,7]

Abstract

Background Monitoring the implementation of new interventions, as in this study Non-violent Resistance (NVR) for the use in residential youth care settings, is mandatory in order to evaluate, adjust and refine the implementation process where necessary.

Objective As there is no instrument for such monitoring of NVR available, the authors developed a new questionnaire, named *Reaction to Unacceptable Behavior Inventory*, (in short: RUBI).

Method This questionnaire was completed by staff of four different residential settings in the Netherlands, at different stages of the NVR implementation process. The staff members reported on the practice of their colleagues.

Results The results are promising, as they show good reliability, inter-item correlations and other psychometric features for the included items. Furthermore, the results show that the RUBI seems to discriminate between trained and untrained teams, defending its use in future implementation processes and implementation research.

Conclusions The RUBI is the first attempt to create an instrument which can be used for monitoring change during implementation of NVR, and for evaluating the degree of difference or compatibility between NVR and existing practice before implementation. Longitudinal research is needed to strengthen the documentation of validity and reliability of the RUBI in different settings, countries, and cultures. This should also be extended to the final and follow-up stages of implementation. In the future, redundant and insensitive items should be removed and standards for interpreting scale scores should be developed.

Keywords Residential youth settings · Questionnaire development · Implementation fidelity · Coping with aggressive behavior

K. M. Visser and K. van Gink have shared first authorship.

✉ K. M. Visser
km.visser@amsterdamumc.nl

✉ K. van Gink
k.vangink@debascule.com

Extended author information available on the last page of the article

Published online: 10 August 2019

Introduction

Any implementation of new interventions in care and treatment depends on whether new practice is actually developing in daily work (Fixsen et al. 2005), and evaluating this should be an integrated part of implementation processes and later service delivery. This requires some method for measuring fidelity at organizational, practitioner or consumer level using self-report or observation. Successful implementation of a certain method requires not only manualisation and effect-evaluations, but also a structured process evaluation, helping to understand why an intervention is successful or not (Burnes 2004; Damschroder et al. 2009; Saunders et al. 2005). As Fixsen et al. (2005, p. 15) state: "Implementation is a process, not an event", which indicates that proper implementation requires sufficient time, money, planning, and personnel resources supporting the implementation (Fixsen et al. 2005). In order to monitor and standardize the different processes of implementation, fidelity assessment can act as a feedback mechanism to improve performance (Breitenstein et al. 2010; Fixsen et al. 2005). Lack of implementation fidelity can weaken outcomes and lead to faulty conclusions about intervention effectiveness. Because they can cause potentially useful interventions to appear ineffective, failures in implementation fidelity are called Type-III Errors (Dobson 2005; Sánchez et al. 2007). To avoid a Type-III-Error, clear and feasible strategies for monitoring and measuring implementation fidelity should be delineated prior to initiation of an intervention study (Breitenstein et al. 2010). This study describes the development and evaluation of a fidelity instrument for Non-violent Resistance (NVR) used by staff in residential treatment institutions. NVR is a methodology to address aggression and other unacceptable behavior in natural and institutional environments. This first version was developed tailored to residential settings.

Since aggression and unacceptable behavior are substantial problems in residential youth settings worldwide (Carlsson et al. 2000; Connor et al. 2004; Knorth et al. 2007; Van der Ploeg 2009; Wielemaker 2009), there is a great need for treatment methods that succeed in reducing this behavior. Over the years, many different approaches have been applied to reduce aggression in residential youth settings. Most of these approaches addressed aggression and unacceptable behavior as an individual problem of the youth (Blake and Hamrin 2007; Connor 2012; Foltz 2004; Fonagy et al. 2002; Kazdin 1987; Lyons and Schaefer 2000; Masters and Bellonci 2002). There is, however, a growing body of evidence showing that aggression can be understood as an interpersonal problem, a complex interaction between patient characteristics, conditions on the ward, and the relationships with staff and parental figures outside the ward (Foltz 2004; Fonagy et al. 2002; Fraser et al. 2016; Nijman et al. 1999).

NVR is a method that addresses aggression as a problem that involves all parties (e.g., the youth, staff and parents). It is a systematic approach to help caregivers, such as parents, teachers, and staff of residential settings, to cope with youths' problematic behavior affecting their environment. NVR builds on restoring the relationship between youth and caretaker, by primarily focusing on changing the behavior of the caretaker instead of wanting to change the behavior of the youth. This is done in a non-violent and non-escalating way. Central tenets of NVR are *presence*, *de-escalation and self-control*, *(re)building relationship*, *support* and *active resistance*. For more detailed explanation of the interventions that are used in NVR, see Omer (2004), Omer and Wiebenga (2015) or Omer and Lebowitz (2016).

NVR has become increasingly popular over the years and more residential settings are now applying this method. NVR appears to be related to a decrease in seclusion and

restraint measures and a more positive atmosphere on the ward, as reported by staff and parents (Goddard et al. 2009). However, despite these documented changes it remains unclear whether and how this results from the implementation of NVR and to what extent NVR has actually become implemented in daily practice. Considering the growing publicity and implementation of NVR in residential settings, an instrument to measure the implementation fidelity is urgently needed, but currently, there are no such instruments available. Research shows that complex interventions (such as NVR) are more difficult to implement with high fidelity, (Craig et al. 2013), and more difficult to evaluate with reliability and validity (Craig et al. 2013). Therefore, developing a measure for implementation fidelity is a challenging but necessary endeavor. This article describes the first steps in the development of such an instrument. As a framework for the development, the guidelines and best practices of instrument development as proposed by DeVellis (2016) were used. The author lays out an eight-step process for developing a scale of questions to measure some construct of the reader's choice. These steps are (slightly reworded): (1) Define clearly what you want to measure, (2) Create a set of draft questions, (3) Select a common format and set of answer options for the questions, (4) Have experts review and revise the questions, (5) Consider using "social desirability" or similar questions, (6) Field test the questions with "real people", (7) Analyze the results of your field test and (8) Decide how many questions–and which questions–to keep (DeVellis 2016).

The aim of this study was to construct an instrument for measuring fidelity to NVR at practitioner level based on participating colleagues observations during the implementation process. The ambition was to develop a first version that may prove usable in implementation monitoring.

The development process used the following strategy and was divided into two parts. As proposed by DeVellis (2016) a combination of an (1) External Expert Evaluation and (2) Psychometric Evaluation was used. After the initial item development with an internal expert group, we expected (1) the external experts to consider each item at least moderately sensitive to NVR practice. Furthermore, we expected the (2) psychometric evaluation to show (a) construct validity of alternative predefined models measured with a Confirmatory Factor Analysis (CFA). The RUBI's psychometric properties were further investigated via examination of the measure's (b) internal consistency for the total score and the predefined subscales/factors. Furthermore, (c) discriminant validity was assessed by known group differentiation and change sensitivity related to NVR implementation: non-trained versus NVR-trained staff members (Hubley 2014) and (d) convergent validity was examined using correlation analysis between the difference in RUBI-NL item scores between non-trained and NVR-trained participants and the item scores on the discriminative measure filled in by the external NVR experts (Hubley 2014). See the Methods section for a more detailed description.

Method

Development of the Instrument

The new instrument was named *Reaction to Unacceptable Behavior Inventory* (RUBI), which was first developed as an English version (RUBI-EN) because English was the shared language between all those invited into the development process. In a later stage of

the development process, a Dutch version (RUBI-NL) was produced (see later), because the psychometric evaluation was planned in Dutch institutions.

Two internal expert meetings were hosted, both consisting of in total eight members. The experts were experienced clinicians, NVR trainers and researchers with prior knowledge about NVR. Before the first meeting, the experts received a first draft of the instruction, item format, and response format for RUBI. This draft did not contain item content, the experts were asked to create potential item content for NVR and other mindsets and behaviors individually. During the first meeting, the eight experts discussed item ideas in subgroups and in a plenary session, which resulted in a large pool of potential items. The experts agreed that the items should be simple and straightforward, and that overly abstract words, double negatives, and ambiguity should be avoided, as suggested by Jebb and Tay (2017). During this meeting some specific challenges were suggested:

First, asking for self-report of practice during training could trigger intended or unintended positive self-portrayal. Asking the informants to describe the mindset and behaviors of their team-members, not their own, should reduce the motivation to portray themselves positively and to avoid social desirability bias (DeVellis 2016). Secondly a more general perspective is that NVR is not an answer to all challenges in residential care and treatment, and may coexist with several other program elements. However, NVR is incompatible with some elements of mindsets and strategies used to correct unwanted behavior. Therefore we chose to contrast core NVR elements with different competing or incompatible principles or strategies, and describe these practices in a positive language to avoid demand characteristics and social desirability influence. It was furthermore decided to scale the responses on a five-point scale with balance between the alternatives in the middle and dominance at either extremes. NVR alternatives were randomly alternated to left or right end of this scale to allowing uncertainty, equivalence, and nuanced response (DeVellis 2016).

Between the first and the second expert meeting, the present authors produced and sent a preliminary draft of the RUBI-EN, based on suggestions from the first meeting, to the experts. They were invited to bring with them suggestions for item removal, item improvement or new items representing missing NVR aspects. During the second session, the eight experts agreed that a final selection of nine mindset and eight behavioral items should be included in the new instrument. Before the development process of RUBI-EN entered its last stage (i.e., the expert and empirical item evaluations), the instrument was translated into Dutch through the back-translation procedure (Brislin 1970). The final version of the questionnaire was sent, both in English and in Dutch, to the four experts with advanced knowledge of both languages, who confirmed translation quality.

Instrument

The RUBI-NL consists of 17 items, including nine items on mindset and eight on behavior. The RUBI-NL should be administered individually, is suitable for repeated measurements, and is intended for (mental health care) professionals working in a residential setting. While filling out the questionnaire, the respondent is instructed to keep in mind a serious and difficult incident that had happened on the ward during the previous 2 weeks. For each item, the respondent can indicate on a five-point scale how likely it is that a certain approach would have been used by the respondents' colleagues. Two example items are presented in Table 1 and the entire questionnaire is freely available and can be obtained via an e-mail to the authors.

Table 1 Example of two items of the RUBI scales mindset and behavior

Mindset						
Colleagues should work together in handling behavioral problems in youth	Mostly A	More often A	Just as often A or B	More often B	Mostly B	Professionals should be able to deal with challenges they meet themselves
Behavior						
We first state the unacceptable behavior, then delay the response in heated situations	Mostly A	More often A	Just as often A or B	More often B	Mostly B	We ignore the event or immediately enforce rules in heated situations

External Expert Evaluation

Participants

Three international NVR experts (a clinician from Israel, working with NVR for a long time, a clinician from the Netherlands working with NVR as a trainer and a researcher working in a NVR-trained environment) known to the authors via a NVR-Network and who had not participated in the development process, were chosen and invited via e-mail to contribute to the development process of the RUBI as external experts.

Procedure

Both, the RUBI-NL and RUBI-ENG, were sent to the external experts and they were asked to rate each item for its ability to differentiate between NVR and other mindsets and behaviors on a five-point Likert scale ranging from 1 (*cannot differentiate)* to 5 (*very good differentiation)* prompted by the question: *"To which degree is this item-pair able to differentiate NVR mindset or NVR behavior from other ways of thinking and behaving in a residential unit?"*. The reason for this expert evaluation was to improve the instrument for its future use, by removing the least sensitive items according to NVR experts. The actual removal was postponed to be based on a combination of expert and psychometric evaluations.

Data-Analysis

First, a content validity index (CVI) was computed for each item (Lynn 1986) based on the ratings of the external experts. As the NVR method does not claim to be fundamentally different from other approaches to cope with aggression and unacceptable behavior, and because NVR is not propagated as contrasting a specific traditional method (Van Gink et al. 2012), the authors of this article chose a rating of 3, 4 or 5 as valid, because they did not expect a radical shift in the entire mindset and behavior of staff. Finally, the agreement between experts regarding items' ability to discriminate NVR and other practices was determined by Fleiss's Kappa coefficient and the intraclass correlation coefficient (ICC). The ICC estimates were calculated using SPSS statistical package version 22.0 (IBM Corp., Armonk, NY) based on a single-rating, consistency, 2-way mixed-effects model.

Psychometric Evaluation

Participants

At four residential youth settings in the Netherlands, 139 staff members were recruited to participate in this study. The sample consisted of 35 males and 104 females with an age range between 21 and 61 years ($M = 34.5$, $SD = 9.5$). Eighteen participants worked at setting A, 34 at setting B, 2 at setting C and 85 at setting D. At the time of the data

collection, participants from setting A were 9 months into the NVR implementation, participants from setting B and C were 6 months into the NVR implementation and participants at setting D had not received NVR training yet.

NVR Training

The NVR training consists of two training days followed by supervision meetings, once every 10 weeks and a follow-up day after 9 months. Attendance is expected of all staff members who are involved in the primary care process (e.g., group workers, parent counsellors, psychiatrists, behavioral therapists, teachers, and managers). The first two NVR training days are a combination of theory and experience-based learning. Staff members learn about the beliefs and mindset central to NVR (e.g., a team must take a stand against unacceptable behavior; preventing aggression is helped by staff members acknowledging and recognizing they have a role in escalation processes; it is an illusion to think one can control other people's behavior; staff and parents need to resist unacceptable behavior from children together). Internalizing these beliefs is a precondition for mastering the NVR attitude. As part of the training, staff members practice the use of NVR communication and practical tools such as delayed response (e.g., strike when the iron is cold), reducing the number of rules, and improving non-verbal and verbal communication skills (e.g., looking at how staff members communicate and how this can lead to escalation of aggression). The staff members get acquainted with the NVR interventions (e.g., the Reparation act, Announcement, and Sit-In) by participating in role-play and, for example, actually practice writing an announcement. The supervision meetings are a mixture of practicing and adjusting interventions, and discussing casuistic or practical dilemmas. The follow-up day is a short recap of theory and the final opportunity to fine-tune the NVR mindset and behavior.

Procedure

Following written informed consent, the RUBI-NL was added to the set of questionnaires, participants filled in every 3 months as part of a longitudinal study measuring the effectiveness of NVR throughout a period of 18 months. The study was approved by the Medical Ethics Committee of the VU medical Centre of Amsterdam (2015.344). The RUBI-NL was administered individually either on paper at the end of a NVR training-day or online after receiving an invitation via email with a personal link to the online version of the questionnaire. Completion of the whole test battery took approximately 35 min.

Data-Analyses

(a) Construct Validity

Confirmatory Factor Analysis (CFA) was used to asses construct validity. Because there are no prior instruments measuring NVR practice and no empirically based factor structure is established, three models based on theory were assessed. The first model was a single-factor model, representing the commonalities of NVR used as baseline for other models. The second model used was a two-factor model based on the distinction between the dimensions (1) Mindset and (2) Behavior as described by Van Gink et al. (2012). The third model was a five-factor model theoretically derived and based on the five central tenets or aspects of NVR (1) Presence, (2) De-escalation, (3) (Re)building the relationship, (4) Resistance and (5) Support (Omer and Lebowitz 2016; Omer

and Wiebenga 2015). The CFA was conducted in R using the package *lavaan*. Robust maximum likelihood (MLR) estimation was preferred over diagonally weighted least squares (WLSMV) estimation, because the RUBI-NL scores were considered continuous and did not follow a multivariate normal distribution. MLR estimation relies less on the assumption of multivariate normal distribution and statistically corrects standard errors and Chi square test statistics (Li 2016). The Satorra-Bentler scaled Chi square was used to test the model, because this statistic is robust in case of a violated normal-distribution assumption (Li 2016). The Chi square statistic was evaluated with caution and supplemented with other fit indices, because the Chi square statistic tends to accept models too often in small sample sizes (Boeije and Boeije 2008; Hu and Bentler 1999). To further assess model fit, the comparative fit index (CFI) and root mean-square error of approximation (RMSEA) with its 90% confidence interval were inspected. As proposed by Hu and Bentler (1999), a good fit between the hypothesized model and the observed data is determined by a CFI value $\geq .95$ and a RMSEA value $< .06$.

(b) Internal Consistency

Internal consistency of the RUBI-NL was determined using Cronbach's alpha, Dillon-Goldstein's rho as proposed by Chin (1998) and finally inter-item correlations, as recommended by Clark and Watson (1995). They proposed a Cronbach's alpha and Dillon-Goldstein's rho of $\geq .70$ and an average inter-item correlation of $\geq .15$ as a criterion for internal consistency.

(c) Discriminant Validity

Because this study is part of a future longitudinal study with stepped wedge design, the participants had received NVR training at different time points, and one group had not yet been trained. Known group validity is a procedure that can demonstrate that a questionnaire can differentiate participants into different groups, based on their score. In this case it was hypothesized that NVR staff who had received 9 months of NVR training would score higher on the RUBI-NL than staff who had received only 6 months of training or no NVR training at all. Furthermore, it was expected that staff who received 6 months of NVR training would score higher than non-trained staff. To test this hypothesis, an ANOVA and several MANOVA's were performed.

(d) Convergent Validity

Convergent validity is commonly tested by measuring a new questionnaire with a previously validated measure of the same construct. As the RUBI-NL is the first questionnaire to asses NVR mindset and behavior, it was decided to perform validity testing against the discriminative measure filled in by experts (described above). In order to assess convergent validity, Spearman's correlation coefficients were used to correlate (1) the difference in RUBI-NL item scores between non-trained and trained participants and (2) the item scores on the discriminative measure filled in by NVR experts. We expected that stronger discriminative items (as valued by the NVR experts) would correlate with items with a larger difference in item score between non-trained and trained staff member. The differences in the RUBI-NL item scores were quantified using standardized effect sizes (SES) calculated as the difference in mean scores between non-trained and NVR trained subgroups divided by the pooled standard deviation of the two subgroups. SES expressed as Cohen's d of 0.2 are considered small, 0.5 moderate, and 0.8 large (Rice and Harris 2005). The score on the discriminative measure filled in by NVR experts, was computed by adding up all expert scores per item. A moderate correlation between the difference in RUBI-NL score and the discriminative score filled in by experts was hypothesized.

Table 2 Descriptive Statistics in non-NVR trained, staff with 6 months training and staff with 9 months training

	N	Mean age (SD)	Female/male	Group worker/staff	
Non-NVR trained staff	85	33.17	(8.9)	63/22	66/19
Staff with 6 months of training	36	33.17	(8.9)	33/3	30/6
Staff with 9 months training	18	37.88	(11.0)	10/8	16/2

Table 3 Confirmatory factor analysis fit indices

	χ^2 (*df*)	CFI	RMSEA (90% CI)
One-factor	147.09 (119)*	.92	.04 (.01–.07)
Two-factor	141.89 (118)	.93	.04 (.00–.06)
Five-factor	141.97 (109)	.93	.04 (.00–.07)
One-factor[a]	112.27 (104)	.97	.02 (.00–.05)
Two-factor[a]	109.82 (103)	.98	.02 (.00–.05)
Five-factor[a]	99.13 (94)	.98	.02 (.00–.05)

*$p < .05$

[a]without *NVR mindset* item 9

Results

Expert Evaluation

According to Lynn (1986), all experts must agree on the content validity of an item (I-CVI of 1.00) if the panel consists of five or fewer experts. After deleting *Mindset 9,* two items still received an item content validity index < 1. However, deleting those items did not lead to a better model fit. Fleiss Kappa was .47 and the intraclass correlation coefficient was .62, which indicates a moderate interrater agreement (Koo and Li 2016; Landis and Koch 1977).

Psychometrical Evaluation

Participant Characteristics

In order to explore whether there were differences between the three groups of employees (non-trained staff, staff after 6 months of NVR training and staff after 9 months of NVR training, a non-parametric Kruskal–Wallis test was used, because the data was non-normally distributed. Staff in the different groups did not differ significantly in age ($\chi^2(2) = 4.38$, $p = .112$) or profession (χ^2 (2) = 1.44, $p = .488$). However, the three groups did differ significantly in male/female ratio $\chi^2(2) = 14.3$, $p = .001$ (see Table 2).

(a) Construct Validity

A one-factor model was used as a baseline comparison against the proposed two-factor and five-factor models. The fit indices are presented in Table 3 and show an adequate fit for the one-factor model. However, the Satorra-Bentler scaled Chi square was significant, which means that the model provided estimates that were significantly different from the observed data. The results further indicate that the two-factor and five-factor model were statistically superior to the one-factor model. However, both models were not satisfying with regard to the model fit (CFI should be above .95). Therefore, we investigated items that could be removed in order to improve model fit based on (a) expert evaluation, (b) modification indices, and (c) inter-item correlations. Item "Mindset 9" (*"Colleagues should not feel alone in handling unacceptable behavior in the youth <–> Professionals should be able to handle unacceptable behavior of the youth independently"*) was the primary candidate according to these criteria, and it can be considered to overlap in content with other items such as "Mindset2" (*"Colleagues should work together in handling behavioral problems in youth").* Deletion led to improved model fit in all three models (see Table 3). Based on these results all three models can be used depending on the need for specification into dimension or aspects. Therefore, item Mindset 9 is omitted from all the following analyses.

(b) Internal Consistency

Cronbach's alpha's were calculated for the total RUBI-NL and for the two- and five-factor models separately. The internal consistency based on Cronbach's alpha of the 16 item RUBI-NL (after deleting item *Mindset* 9) was good ($\alpha = .79$). Composite construct reliability, based on Dillon-Goldstein's rho ($\rho = .84$) and the internal consistency as indicated by the average inter-item correlation were acceptable ($r = .20$). Reliability for the subscale *Mindset* was low ($\alpha = .59$) based on Cronbach's alpha, but good based on Dillon-Goldstein's rho ($\rho = .74$) and the reliability for the subscale *Behavior* was acceptable ($\alpha = .72$) to good ($\rho = .81$). The internal consistency as indicated by the average inter-item correlation for the subscales *Mindset* ($r = .16$) and for Behavior ($r = .25$) were both acceptable.

Reliability for the different NVR aspects (i.e., the five-factor model) were low, ($\alpha = .11$, *presence*; $\alpha = .44$, *de-escalation*; $\alpha = .59$ *relationship;* $\alpha = .64$ *resistance* and $\alpha = .30$ *support)* based on Cronbach's alpha. Composite construct reliability was good for *de-escalation, relationship and resistance* ($\rho > .70$) except for *presence* and *support*. The average inter-item correlation was acceptable, except for *presence* ($r = .12$). Because of the low reliability values we decided to turn down the evaluation of the five-factor model and only use the on-factor and two-factor models.

(c) Discriminant Validity

Table 4 presents the means, standard deviations, F values and the results of post hoc comparisons for the three groups differing in implementation stage on the total RUBI-NL and the subscales (mindset and behavior)). A one-way ANOVA for the total scale and MANOVA's for each subscale revealed significant group differences on the total and dimensional scales (behavior and mindset). To summarize, the results generally support the known groups validity of the RUBI-NL total score and the dimensional subscales (behavior and mindset).

(d) Convergent Validity

Results of the Spearman correlation indicated that there was a significant positive association between the RUBI-NL item and the expert discriminative item score, ($rs = .57$, $p = .021$).

Table 4 Known-groups validity: comparisons of group means on the RUBI-NL total score and subscales

	Non-trained staff (n = 85)		6 months of training (n = 36)		9 months of training (n = 18)		F	df	Post hoc comparisons ($p < .05$)
	M	SD	M	SD	M	SD			
1 factor model									
Total	3.47	0.43	3.54	0.48	3.89	0.46	6.55**	2136	0 < 9 6 < 9
2 factor model									
Mindset	3.47	0.47	3.56,	0.48	3.88	0.41	5.50**	2136	0 < 9
Behavior	3.47	0.52	3.52	0.58	3.91	0.55	5.00**	2136	0 < 9 6 < 9

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

Discussion

The aim of the present study was to provide researchers and practitioners with a first version of a psychometrically validated instrument to measure implementation fidelity of Non-violent Resistance (NVR) for use in residential settings. The Reaction to Unacceptable Behavior-Inventory (RUBI-NL) was constructed during expert meetings. Content validity was evaluated by experts who had not contributed to the development of the instrument. Based on their evaluation, it was concluded that the 16-item instrument appeared to have good content validity and that it differentiated between NVR and other methods. Furthermore, construct validity was assessed using confirmatory factor analysis. After deletion of one item, the results indicated the two-factor model with the subscales *Mindset* and *Behavior* as the best model. The one- factor model met the criteria for acceptance as well. Additionally, the RUBI-NL seemed reliable and valid based on the high correlations between the items and the moderate to good internal consistency for the one-factor and two–factor models, after the removal of one item. The five-factor model (with the different NVR aspects) seemed less reliable with regard to its internal consistency. Lastly, the RUBI-NL appeared to discriminate between NVR-trained staff and non-trained staff, in case of total NVR score, NVR Mindset and NVR Behavior. This discrimination was between non-trained staff and staff with 9 months of training. Staff with 9 months of NVR training scored significantly higher than non-trained staff. A possible explanation for the fact that there was almost no difference between non-trained staff and staff members with 6 months of training could be that this method is more than a change in ways but aims at a change in culture, what could possibly take more time than 6 months of training.

Taken together, the results suggest that this instrument seems reliable and valid to measure compatibility and progress in implementation of the NVR method in residential settings, and that it can be scored as a total score, as well as subdivided in two dimensions (Mindset and Behavior). This validation study was conducted with staff members in residential settings with no-, or partial implementation of the NVR method. As a whole, the RUBI-NL has provided evidence to be reliable for preliminary use. Further studies including method survival follow-up to complete the validation and establish reference values for use in evaluation of compatibility as well as implementation progress and fidelity are needed.

The external NVR expert evaluations area weak point in this study. It was challenging to find ways to differentiate NVR from other practices, which increased the risk of achieving

only moderate differentiating power. In addition, asking only three experts to evaluate the discriminative power of items, made our tests vulnerable to individual variations between the experts, a concern which is confirmed by only moderate inter-rater agreement between the external experts. The internal experts participating in the instrument development all agreed on the choice of items. Therefore, we chose not to exclude items based on only external expert evaluations. The item suggested for deletion by CFA modification indices and contribution to internal consistency, was also below the desired level of external expert evaluation. Two other items, which also scored below this level of external expert evaluation, were kept in the final version because there was no other reason to remove them (e.g. based on CFA modification indices or contribution to internal consistency). Possibly, items could be changed and rearranged in a future validation process.

Lastly, the relatively small sample size of < 150 participants might have affected the results, as with smaller sample sizes findings are not always generalizable. As respondents stemmed from four different institutions and varied in gender, age and work experience, auteurs tried to circumvent this limitation. In a future study, sample size should be larger in order to draw more solid conclusions.

Conclusion

The RUBI-NL was developed to evaluate implementation of the NVR-method with repeated measurements during the implementation process. Knowledge on the status of implementation is important because lack of or incomplete implementation can lead to false conclusions about intervention effectiveness and type-II Errors (Dobson 2005; Sánchez et al. 2007).

Despite the relatively small sample size and a limited range of institutions and implementation stages, the study has documented promising psychometric qualities and discrimination between trained and untrained teams, defending its use in future implementation and implementation research. Longitudinal research is needed to strengthen documentation of validity and reliability of the RUBI-NL in different settings, countries and cultures. This should also be extended to the final and follow-up stages of implementation. Finally, standards for interpreting scale scores should be developed.

Although further research is necessary to test and develop this promising instrument, it can be used to ensure implementation quality and longitudinal fidelity. The authors will make the RUBI free and openly available in Dutch, English, and Norwegian, only requesting data sharing in exchange.

Acknowledgements The authors would like to thank all staff members of the three child and adolescent residential settings that were willing to provide information for this paper.

Funding This study was funded by Mind Netherlands (2013 6746); Ministerie van Volksgezondheid, Welzijn en Sport (57415); Netwerk Effectief Jeugdwerkstelsel Amsterdam.

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed Consent Informed consent was obtained from all individual participants included in the study.

References

Blake, C. S., & Hamrin, V. (2007). Current approaches to the assessment and management of anger and aggression in youth: A review. *Journal of Child and Adolescent Psychiatric Nursing, 20*(4), 209–221.

Boeije, H., & Boeije, H. (2008). *Analyseren in kwalitatief onderzoek*. Amsterdam: Boom onderwijs.

Breitenstein, S. M., Gross, D., Garvey, C. A., Hill, C., Fogg, L., & Resnick, B. (2010). Implementation fidelity in community-based interventions. *Research in Nursing & Health, 33*(2), 164–173.

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology, 1*(3), 185–216.

Burnes, B. (2004). Emergent change and planned change–competitors or allies? The case of XYZ construction. *International Journal of Operations & Production Management, 24*(9), 886–902.

Carlsson, G., Dahlberg, K., & Drew, N. (2000). Encountering violence and aggression in mental health nursing: A phenomenological study of tacit caring knowledge. *Issues in Mental Health Nursing, 21*(5), 533–545.

Chin, W. W. (1998). The partial least squares approach to structural equation modeling. *Modern Methods for Business Research, 295*(2), 295–336.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*(3), 309.

Connor, D. F. (2012). *Aggression and antisocial behavior in children and adolescents: Research and treatment*. New York: Guilford Press.

Connor, D. F., Doerfler, L. A., Toscano, P. F., Volungis, A. M., & Steingard, R. J. (2004). Characteristics of children and adolescents admitted to a residential treatment center. *Journal of Child and Family Studies, 13*(4), 497–510.

Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I., & Petticrew, M. (2013). Developing and evaluating complex interventions: The new Medical Research Council guidance. *International Journal of Nursing Studies, 50*(5), 587–592.

Damschroder, L. J., Aron, D. C., Keith, R. E., Kirsh, S. R., Alexander, J. A., & Lowery, J. C. (2009). Fostering implementation of health services research findings into practice: A consolidated framework for advancing implementation science. *Implementation Science, 4*(1), 50.

DeVellis, R. F. (2016). *Scale development: Theory and applications* (Vol. 26). Beverley Hills: Sage Publications.

Dobson, K. S. (2005). Definitional and practical issues in the assessment of treatment integrity. *Clinical Psychology: Science and Practice, 12,* 384–387.

Fixsen, D. L., Naoom, S. F., Blase, K. A., & Friedman, R. M. (2005). *Implementation research: A synthesis of the literature*. Tampa: the National Implementation Research NetworkLouis de la Parte Florida Mental Health Institute, University of South Florida.

Foltz, R. (2004). The efficacy of residential treatment: An overview of the evidence. *Residential Treatment for Children & Youth, 22*(2), 1–19. https://doi.org/10.1300/J007v22n02_01.

Fonagy, P., Target, M., Cottrell, D., Phillips, J., & Kurtz, Z. (2002). *What works for whom? A critical review of treatments for children and adolescents* (Vol. 2). New York: Guilford Press.

Fraser, S. L., Archambault, I., & Parent, V. (2016). Staff intervention and youth behaviors in a child welfare residence. *Journal of Child and Family Studies, 25*(4), 1188–1199.

Goddard, N., Van Gink, K., Van der Stegen, B., Van Driel, J., & Cohen, A. (2009). Smeed het ijzer als het koud is. *Non-violent resistance op een acuut psychiatrische afdeling voor adolescenten. Maandblad Geestelijke Volksgezondheid, 64,* 531–539.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55.

Hubley, A. M. (2014). Discriminant validity. In A. C. Michalos (Ed.), *Encyclopedia of quality of life and well-being research* (pp. 1664–1667). Dordrecht: Springer.
Jebb, A. T., & Tay, L. (2017). Introduction to time series analysis for organizational research: Methods for longitudinal analyses. *Organizational Research Methods, 20*(1), 61–94.
Kazdin, A. E. (1987). Treatment of antisocial behavior in children: Current status and future directions. *Psychological Bulletin, 102*(2), 187.
Knorth, E. J., Klomp, M., Van den Bergh, P. M., & Noom, M. J. (2007). Aggressive adolescents in residential care: A selective review of treatment requirements and models. *Adolescence, 42*(167), 461.
Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163.
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174.
Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods, 48*(3), 936–949.
Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research, 35,* 382–385.
Lyons, J. S., & Schaefer, K. (2000). Mental health and dangerousness: Characteristics and outcomes of children and adolescents in residential placements. *Journal of Child and Family Studies, 9*(1), 67–73.
Masters, K. J., & Bellonci, C. (2002). Practice parameter for the prevention and management of aggressive behavior in child and adolescent psychiatric institutions, with special reference to seclusion and restraint. *Journal of the American Academy of Child and Adolescent Psychiatry, 41*(2), 4S–25S.
Nijman, H. L., á Campo, J. M., Ravelli, D. P., & Merckelbach, H. L. (1999). A tentative model of aggression on inpatient psychiatric wards. *Psychiatric Services, 50*(6), 832–834.
Omer, H. (2004). *Non-violent resistance: A new approach to violent and self-destructive children*. New York: Cambridge University Press.
Omer, H., & Lebowitz, E. R. (2016). Nonviolent resistance: Helping caregivers reduce problematic behaviors in children and adolescents. *Journal of Marital and Family Therapy, 42*(4), 688–700.
Omer, H., & Wiebenga, E. (2015). *Geweldloos verzet in gezinnen: een nieuwe benadering van gewelddadig en zelfdestructief gedrag van kinderen en adolescenten*. Berlin: Springer.
Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. *Law and Human Behavior, 29*(5), 615–620.
Sánchez, V., Steckler, A., Nitirat, P., Hallfors, D., Cho, H., & Brodish, P. (2007). Fidelity of implementation in a treatment effectiveness trial of reconnecting youth. *Health Education Research, 22*(1), 95–107. https://doi.org/10.1093/her/cyl052.
Saunders, R. P., Evans, M. H., & Joshi, P. (2005). Developing a process-evaluation plan for assessing health promotion program implementation: A how-to guide. *Health Promotion Practice, 6*(2), 134–147.
Van der Ploeg, J. (2009). *Agressie*. Rotterdam: Lemniscaat.
Van Gink, K., Van der Stegen, B., Goddard, N., & Ottenbros, R. (2012). *Non-violent Resistance in de (semi) residentiële setting. Een nieuwe aanpak van agressief en destructief gedrag voor teams*. Amsterdam: de Bascule.
Wielemaker, J. (2009). *Langdurig klinisch behandeld in de kinder-en jeugdpsychiatrie; een follow-up onderzoek na 5 tot 25 jaar [Long-term residential treatment in child and adolescent psychiatry; a follow-up study after 5 to 25 years]*. Unpublished doctoral dissertation, Erasmus University, Rotterdam.

Affiliations

K. M. Visser[1] · K. van Gink[1,2] · F. Thissen[3] · T. A. Visser[4] · T. Rimehaug[5,6] · L. C. M. Jansen[1] · A. Popma[1,7]

1 Department of Child and Adolescent Psychiatry, VU University Medical Center, Meibergdreef 5, 1105 AZ Amsterdam, The Netherlands

2 de Bascule, Academic Center for Child and Adolescent Psychiatry, Meibergdreef 5, 1105 AZ Amsterdam, The Netherlands

3 Department of Developmental and Educational Psychology, The Faculty of Social and Behavioral Sciences, Leiden, The Netherlands

4 Department of Psychology, Education and Child Studies, Erasmus School of Social and Behavioural Sciences, Rotterdam, The Netherlands

5 Regional Centre for Child and Youth Mental Health and Child Welfare-Central Norway, Department of Mental Health, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

6 Department of Child and Adolescent Psychiatry, Nord-Trøndelag Health Trust, Levanger, Norway

7 Department of Criminal Law and Criminology, Leiden University, Leiden, The Netherlands