Evaluation and Program Planning 99 (2023) 102299

Contents lists available at ScienceDirect

Evaluation and Program Planning

journal homepage: www.elsevier.com/locate/evalprogplan

Development and evaluation of a fidelity tool in a post-discharge responsive parenting intervention program for very preterm born children

Monique Flierman [a,*], Eline Vriend [a], Aleid G. Leemhuis [c], Raoul H.H. Engelbert [a,b,c], Martine Jeukens-Visser [a]

[a] Amsterdam UMC, University of Amsterdam, Department of Rehabilitation, Amsterdam Reproduction and development, Meibergdreef 9, Amsterdam, the Netherlands
[b] Centre of Expertise Urban Vitality, Faculty of Health, Amsterdam University of Applied Sciences, Amsterdam, the Netherlands
[c] Emma Children's Hospital, Department of Pediatrics, Amsterdam UMC, University of Amsterdam, Meibergdreef 9, Amsterdam, Netherlands

ARTICLE INFO

Keywords:
Fidelity tool
Implementation
Parent-child interaction
Very preterm birth
Early intervention

ABSTRACT

The TOP program is a fully implemented responsive parenting intervention for very preterm born infants. Fidelity monitoring of interventions is important for preserving program adherence, impact outcomes and to make evidence-based adaptations. The aim of this study was to develop a fidelity tool for the TOP program following an iterative and co-creative process and subsequently evaluate the reliability of the tool. Three consecutive phases were carried out. Phase I: Initial development and pilot testing two methods namely self-report and video based observation. Phase II: Adaptations and refinements. Phase III: Evaluation of the psychometric properties of the tool based on 20 intervention videos rated by three experts.The interrater reliability of the adherence and competence subscales was good (ICC.81 to .84) and varied from moderate to excellent for specific items (ICC between .51 and .98). The FITT displayed a high correlation (Spearman's rho.79 to.82) between the subscales and total impression item. The co-creative and iterative process resulted in a clinical useful and reliable tool for evaluating fidelity in the TOP program. This study offers insights in the practical steps in the development of a fidelity assessment tool which can be used by other intervention developers.

1. Introduction

Despite widespread agreement on the need for post-discharge support for very preterm (VPT) infants and their families, it is not often integrated into routine post-discharge clinical practice (Anderson, Treyvaud, & Spittle, 2020). In the Netherlands, a preventive home-based intervention program (the TOP program) for VPT children (< 32 weeks of gestational age and/or birth weight < 1500 g) and their parents is fully implemented, reimbursed by all Dutch Health insurance companies, and reaches yearly more than 75% of all Dutch VPT children (Jeukens-Visser et al., 2020) (Perined, 2020). The TOP program is a process-oriented intervention with a theoretical framework that includes seven key strategies to target the desired outcomes (Fig. 1; Theory of Change). The program is based on extensive research with sustained positive effects on infant cognitive, motor, and behavioral outcomes (Koldewijn et al., 2010; Van Hus et al., 2016) and has been gradually implemented (Jeukens-Visser et al., 2020)

Scaling up an intervention from the controlled research conditions to routine care is long and complex process and requires selection of strategies relevant and feasible to that specific context (Fixsen, Blase, Naoom, & Wallace, 2009). An important aspect during this journey is implementation fidelity; the degree to which the intervention is delivered as intended. (Cross and West, 2011) Low fidelity could diminish the effect of interventions, but adaptations that suit the needs and preferences of patients may also improve effectiveness. Although fidelity is associated with positive outcomes, it may be difficult to achieve in routine care and sustaining program fidelity of large-scale disseminated interventions has rarely been studied systematically (Askeland, Forgatch, Apeland, Reer & Gronlie, 2019; Durlak & DuPre, 2008, Huth--Bocks, Jester Stacks, Muzik Rosenblum & Michigan, 2020).

Fidelity assessments are especially important to use for interventions where multiple factors can influence the fidelity of the intervention delivery, such as heterogeneity of interventionists, participant characteristics, and family circumstances (Tiddmarsh, Whiting, Thompson &

Abbreviations: TOP program, Transmural developmental support for VPT infants and their parents; FITT, Fidelity of Implementation Tool TOP program; TOC, Theory of Change; VPT, Very preterm.

* Correspondence to: Amsterdam UMC, University of Amsterdam, Department of Rehabilitation, Meibergdreef 9, Amsterdam, the Netherlands.
E-mail address: m.flierman@amsterdamumc.nl (M. Flierman).

https://doi.org/10.1016/j.evalprogplan.2023.102299
Received 27 January 2022; Received in revised form 17 March 2023; Accepted 29 April 2023
Available online 4 May 2023
0149-7189/© 2023 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Cumming, 2022). In order to evaluate the fidelity of the TOP program we first needed to develop an intervention-specific fidelity assessment tool. Despite the presence of validated fidelity tools for several parent-child interventions, these can only serve as examples as the target groups and the intervention strategies differ from the TOP program. (Song et al., 2010; An et al., 2021; Bastick et al., 2018; Breitenstein, Fogg, Garvey, Hill, Resnick, & Gross, 2010; Di Rezze, Law, Eva, Pollock, & Gorter, 2013; Forgatch & DeGarmo, 2011; Goering et al., 2016; Parham et al., 2011). There is a great diversity in concepts and constructs for establishing fidelity and fidelity measures used for parenting programs differ from observational methods such as live observations, video-recorded sessions and non-observational measures such as self-report. (Gearing et al., 2011; Martin, Steele, Lachman & Gardner, 2021). An (2020) described a usefull multidimensional construct for intervention fidelity, including adherence to key components, quality of intervention delivery, amount of intervention delivered, participant responsiveness, and program differentiation. However, guidance on the processes involved in developing a feasible and valid fidelity tool within complex interventions at scale is limited (Toomey, Matthews, Guerin & Hurley, 2016).

The TOP program is part of a interdisciplinary learning community created by the Center on the Developing Child at Harvard University. Their IDEAS Impact Framework™ provides an innovative approach for program development and evaluation, in order to improve outcomes. (Center on the Developing Child; IDEAS impact framework, 2015) During the implementation of the TOP program, the four guiding principles: precision, fast-cycle iteration, co-creation, and shared learning, have been adopted (Schindler, Fisher & Shonkoff, 2017). Starting with the precision, a Theory of Change of the TOP program was developed, a framework that links interventions strategies, via targets, to the outcomes. The next step is to include a fidelity measure that is closely tied to our intervention strategies and training materials. Following an iterative process enables rapid learning and making refinements while developing the tool. A partnership between researchers, developers, educators and interventionists contributes to the development of a realistic and practicable tool. This increases the likelihood that the tool has both content validity and is feasible for use in practise.

This study aimed to develop an intervention-specific fidelity tool, in co-creation with stakeholders, and study the psychometric properties to be able to guarantee the quality of the scaled-up TOP program.

2. Methods

This study consisted of three distinct phases. In phase I, a draft of a fidelity tool was developed and tested in a co-creation group. The co-creation group consisted of five certified TOP interventionists with more than two years of experience in the execution of the TOP program and three members of the research team. The results of the pilot testing in phase I gave rise to an extensive evaluation and necessary adaptations. In phase II, the co-creation group evaluated in two work sessions the results of the pilot testing and user experiences. Recommendations for improvements were formulated and concerned the rating method, the item content, and training of expert raters. The adaptations led to the Fidelity of Implementation Tool TOP program (FITT). In phase III, the psychometric properties of the FITT were studied.

The Medical Ethical Review Committee (METC) of the Academic Medical Center Amsterdam provided a waiver for ethical review of this research. Parents signed informed consent for collecting personal data, video recording, and self-report evaluations following the Privacy Act. In addition, the TOP interventionists signed informed consent for being recorded.

2.1. Participants

Families who participated in the TOP program with a VPT child (< 32 weeks of gestation) or with very low birth weight (< 1500 g) were eligible for both the pilot- and psychometric evaluation study. Socio-demographic and perinatal factors of included participants are presented in Table 1.

We classified the children in an even distribution of child's corrected age (CA) since the deployment of intervention strategies could differ per developmental stage and during the course of the 1-year TOP program.

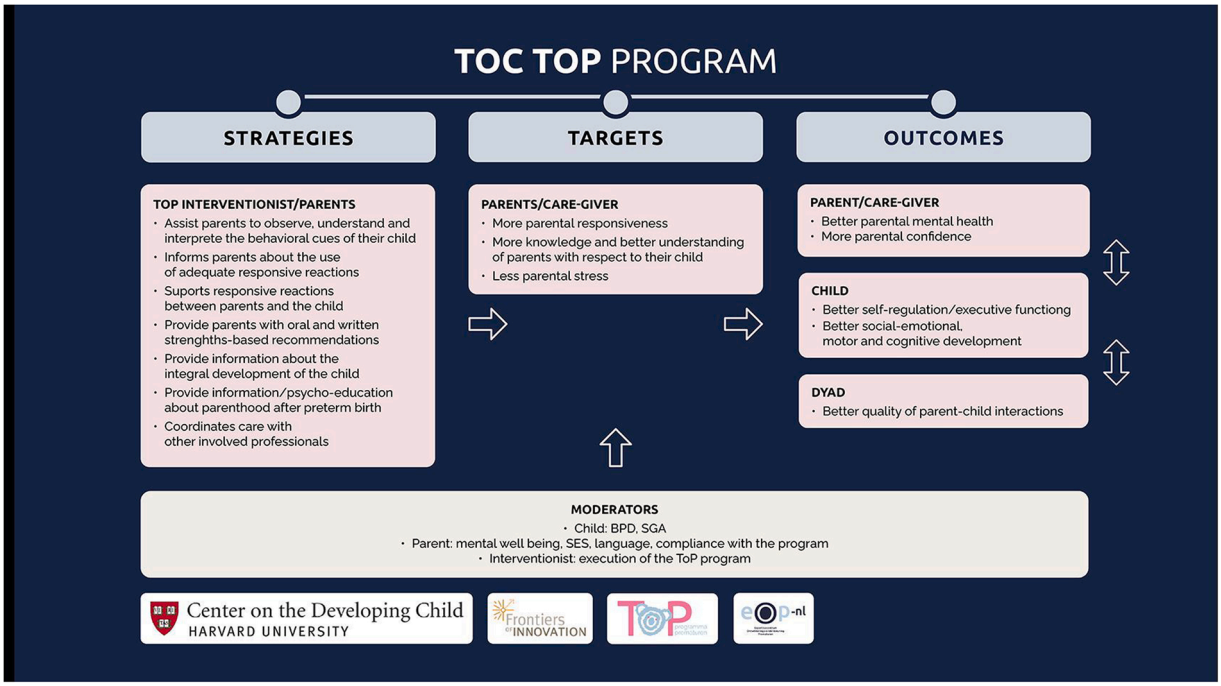

Fig. 1. Theory of Change of the TOP program.

Table 1
Participant characteristics.

	Total group (N = 19 families)
Perinatal factors	
Gestational age, *n*	2
	15
• < 28 weeks	2
• 28–32 weeks	
• ≥ 32 weeks	
Birth weight (gram), *M*(*SD*)	1419 (344)
Days in hospital, *M*(*SD*)	53 (23)
Gender child boy, *n*	11
Singleton, *n*	19
Age during the home visit (CA), *n*	3
	9
• 0–3 months	8
• 3–6 months	
• > 6 months	
Social factors	
Age mother, years, *M*(*SD*)	34 (5.8)
Age father, years *M(SD)*	35 (7.5)
Mother born in the Netherlands, *n*	15
Father born in the Netherlands, *n*	16
Dutch speaking family, *n*	16
Marital state married/together, *n*	17
Level of education - Mother, *n*	1
	9
• Low	9
• Middle	
• High	
Level of education - Father, *n*	1
	8
• Low	9
• Middle	1
• High	
• Missing	
No. home visit – median, range	5.5 (3–11)
Number of visits with both parents, *n*	4

There were 20 videos from 19 families.
[b]Low = primary school, pre-vocational secondary education; middle = senior general secondary, education, pre-university, or secondary vocational education; high = higher professional education or university education.

2.2. Tool development

2.2.1. Phase I: Development and pilot testing

The first author (MF) developed a preliminary tool based of two scales; 1) adherence to program content and 2) competence in program delivery. Key intervention strategies from the Theory of Change (TOC) (Fig. 1) were operationalized into observable items for the Adherence scale. The techniques to enhance the execution of the key TOP strategies were operationalized for the Competence scale. Two key strategies of the TOP program ('written strength-based recommendations' and 'co-ordinates care with other involved professionals') took place outside the home visit and were therefore not included in the tool. Ratings were based on the number of opportunities the interventionist utilized to apply the described key strategy or used a competence technique. At the same time, missed opportunities led to a lower score on that item. Two methods were explored for measuring the fidelity, namely a self-monitoring- and a video-based observation tool. The preliminary draft of the measuring tools was discussed, clarified, and refined in the first co-creation session. Then, video-recorded intervention sessions were used to discuss the items and response options critically. In between sessions, the TOP interventionists reflected on how well the FITT draft matched their actual intervention execution after their regular TOP home visits. In the following co-creation sessions, all items were further discussed to determine if they should be retained, modified, or missed core components.

The adherence scale was operationalized in eleven items that measured the frequency of applying the described key strategies, on a 3-point rating scale. For the competence scale, four items that represented the strength-based approach and skills to transfer information to the parents were identified, rated on a 3-point scale. A total impression item was added to assess the overall quality of the intervention and was scored on a 10 point Likert scale with a score ranging from 1 (very low) to 10 (excellent). In addition, the therapists indicated the need to outline family characteristics that may influence parental engagement. Therefore, the item "parental availability" was added, measuring how receptive or engaged the parent was during the intervention session, also rated on a 3-point scale. The manual, including detailed descriptions and illustrative examples, guidelines for scoring, and decision rules, was created simultaneously.

2.3. Procedures Pilot testing

TOP interventionists (n = 5) were video recorded during two of their routine TOP home visits. In addition, they filled in the self report after ten of their TOP home visits, including the two video-recorded home visits. An independent researcher (EV) compressed the recorded home visit (1 h) to capture the intervention's core in about 25 min. The five TOP interventionists rated the video-recorded home visits (n = 8), except their own, to assess the inter-rater reliability.The first author (MF) rated the complete video-recorded home visits. The self-reported scoring of the raters was compared with the rating of the first author.

2.3.1. Phase II: Adaptations and refinement

In two work sessions, the co-creation group evaluated the results of the pilot testing and user experiences. Regarding the video-based rating, all individual items were discussed using the comments and reflection notes from the participating TOP interventionists. In addition, the rating system was discussed, and the reliability of the video rating by TOP interventionists as raters was evaluated. The research team changed and clarified definitions per item and tested the adaptations with training videos.

2.3.2. Phase III: Psychometric evaluation of the FITT

2.3.2.1. The FITT instrument. The FITT contains four items for the adherence scale and five items for the competence scale, a global impression score (item 10), and an item on parental availability (item 11.1 and 11.2). Items 1–9 and 11 (parental involvement during the home visit) are rated on a 3-point Likert scale. Each score per item is described comprehensively and contains examples (See Inline supplementary Appendix S2), and scores 1–3 indicate ascending quality of the execution. Item 10 is the global impression score on a scale of 1 (very bad impression) to 10 (excellent impression).

Procedures.

Three TOP lecturers, assigned as expert raters, evaluated twenty regular TOP home visits videos. The collected videos were reviewed for sufficient quality and diversity by one researcher (EV). Just as in phase I, complete video recorded home sessions were compressed into 20–30 min videos to capture the intervention's core. The manual with guidelines for the use of the FITT, description of the FITT items with examples and scoring criteria accompanied the video vignets. Rating of the items was performed after the first observation of the video vignette, and the final scoring was given after the second viewing.

2.4. Statistical analyses

A dataset was created in Castor, a web-based system to build electronic Case Report Forms (CRF). Subsequently, the data was exported to Statistical Package for the Social Sciences (SPSS, version 26.0).

In phase I, the agreement between self-report and expert video rating for the composite scores of the subscales were measured with a two-way random effects model, single measures. (Shrout & Fleiss, 1979). Cohen's

Kappa with quadratic weighting (Fleiss & Cohen, 1973) was calculated for individual items of the FITT. Interrater reliability of the video ratings was measured with Krippendorf's alpha (Kalpha). The Kalpha was chosen because each video had missing values, as the TOP interventionist did not rate their intervention sessions. Furthermore, the Kalpha does not require minimal sample sizes (Hayes & Krippendorff, 2007).

In phase III, the inter-rater reliability of the subscales and the individual items was measured with the ICC (two-way mixed model, mean-rating, absolute agreement) with a 95% confidence interval. Values of the ICC were described as poor (<0.50), moderate (0.50 −0.75), good (0.75 −0.90) or excellent (>0.90) (Koo & Li, 2016) Chronbach's alpha (α) was used to measure the internal consistency of the subscales adherence and competence. The Spearman's Rho (r_s) was calculated to examine correlations between the adherence and competence scales and both to the total impression score.

3. Results

3.1. Phase I

The agreement between the self-report (n = 10) from the five interventionists and expert video rating was below the acceptable level of agreement (ICC<0.5) for individual items and the composite scales. The inter-rater reliability for individual items measured with the Kalpha was poor for the individual items and the two subscales.

3.2. Phase II

The evaluation led to necessary adaptations (See Inline supplementary Appendix S1). Recommendations for improvements were formulated and concerned the (1) method of rating, (2) item content, and (3) raters.

All five TOP interventionists agreed that recalling the details of the home visit was too difficult. Since the rating scale was based on the number of used opportunities, the lack of awareness or recalling missed opportunities led to possible over-estimation of the used strategies. The context of the session, family circumstance, and prior interventions sessions played a role when filling out the self-report, also causing subjective and more positive assessments by the interventionist. Therefore, the development of the self-report fidelity tool was stopped. Furthermore, instead of rating the number of used opportunities, it was decided to provide all items with specific objective criteria to represent the particular level of quality Items were scored on a Likert scale (0 = No or insufficient execution of the strategy/competence, 1 = sufficient execution of the strategy/competence, 2 = good execution of the strategy/competence. Additional guidelines, explanations, and examples for scoring per item were added to the manual.

Regarding item content, changes and clarifying definitions per item were made, and some competence items were added and adapted to improve the competence scale of the tool further (See Inline supplementary Appendix S1). Finally, the video rating by interventionists as raters was evaluated. Although the raters worked closely on developing the scale, deviant interpretation of the items and no additional training in the use of the tool may have caused the poor inter-rater reliability. Another aspect, rating fellow co-workers could have counted for higher ratings. Extensive training of fidelity raters in this phase of the iterative and flexible development process was not considered efficient. Therefore, the three experienced lecturers with additional training in rating video's were assigned as expert-raters for phase III.

3.3. Phase III

Interrater reliability from the three expert raters across the 20 videos for both subscales and total impression score were excellent, with average measures ICC > 0.80 [95% CI.60 to.95, p < .001] (Table 2).

For the individual items, the inter-rater reliability was substantial; nine items scored ICC values > 0.60 [95% CI, range 0.62-0.99]. Only item 4 (Informing about development and parenting) of the adherence scale and item 9 (Use of didactic skills) of the competence scale had ICC scores < 0.60 [range.51 to.56] (Table 2). The internal consistency as measured by Chronbach's alpha (α) was .57 for the adherence scale and .77 for the competence scale. The associations between the subscales and each subscale to the total impression scale were (very) strong with Spearman's rho ranging between r_s0.70 and 0.82.

Table 2
Results interrater reliability of the FITT (item and composite scores).

	ICC	95% CI
Item 1 – reading behavioral cues	0.645	0.271 – 0.846
Item 2 – promoting adequate responsive reactions	0.852	0.690 – 0.937
Item 3 – creating conditions for enhancing development	0.617	0.062 – 0.779
Item 4 – informing about development and parenting	0.510	0.122 – 0.822
Item 5 – discussing intervention goals	0.804	0.588 – 0.916
Item 6 – intervening in parent-child interaction	0.720	0.422 – 0.885
Item 7 – using strength-based approach	0.635	0.258 – 0.841
Item 8 – timing and dosing in information transfer	0.811	0.607 – 0.919
Item 9 – use of didactic skills	0.562	0.127 – 0.808
Item 11.1 – availability of the mother/parent 1	0.616	0.213 – 0.834
Item 11.2 – availability of the father/parent 2	0.986	0.971 – 0.994
Composite scores and global impression		
Adherence scale	0.808	0.601 – 0.917
Competence scale	0.842	0.666 – 0.933
Total impression score	0.877	0.764 – 0.952

Note. ICC estimates were based on a mean-rating (k = 3), absolute-agreement, two-way mixed-effects model; Values of the ICC were described as poor (<0.50), moderate (0.50 −0.75), good (0.75 −0.90), or excellent (>0.90)

4. Discussion

This study describes the process of developing and evaluating an intervention-specific fidelity measure for the implemented TOP program for VPT and their parents. The co-creative and iterative process led to a reliable tool that objectifies intervention fidelity and is relevant for adding transparency and transferability to the TOP program.

4.1. Self report vs. video based observation

Initial evaluation of the reliability in phase I was not very promising. The results showed subjective bias in the self-report: higher fidelity ratings compared to the observational measure. The interventionists agreed that scoring retrospectively and objectively recalling details with the fidelity checklist was too difficult. Although this systematic bias in fidelity self-reports is confirmed by other studies, the self-report method would have been less complicated and costly in an integrated routine post-discharge intervention and was therefore explored (Breitenstein et al., 2010; Mowbray, 2003, Tidmarsh, Whiting, Thompson, & Cumming, 2022).

The drawback of the video observations with high labor costs, time involved to train raters and the challenges to recruit families and interventionists outweighed the rich and unique source of information coming from video observations of regular TOP home visits. The results of this study support the use of videotaped interventions as a meaningfull method for fidelity assessment (Asan & Montague, 2014; Cross et al., 2015).

4.2. Translating the key strategies and competent execution into observable items

The TOP program is a complex intervention since it contains several interacting components and interventionist have to tailor the intervention to the family context to carry out the key strategies of the program. We wanted to assure that the fidelity tool captured the adherence to the key strategies as described in the Theory of Change (fig1) and the competence in program delivery following the described pedagogical

approach and vision of the program. However, the more detailed and subdivided elaboration of the key strategies in phase I led to substantive interpretation issues. The level of detail was reduced and described in more observable behavior to leave less room for ambiquity in interpretation and scoring. The amount of items representing the key strategies was reduced to make the tool applicable for future purposes.

Since each item in the FITT represents a different key strategy or meaningful competence, we aimed for ICC values > 0.50 at the item level. Two items (4, 9) only just complied with the ICC criteria and might need better descriptions and examples.The interrater reliability of the Adherence and Competence scales and the total impression scale scored in the excellent range. Composed scores usually demonstrate stronger reliability than individual items (Parham et al., 2011). The good agreement between the subscales and each subscale to the total impression scale indicates that these FITT scores can be used to establish insight in the intervention fidelity. By relying on a single video observation to rate the fidelity of a 1 year process-oriented intervention we could risk making erroneous conclusions on the fidelity, as also decribed by Cross et al., 2015. Which number of home visits should be rated with the FITT to obtain insight into the overall quality of the execution of the intervention needs to be determined in future research.

4.3. Process of development

We partnered with TOP interventionists, researchers and program developers to develop a fidelity monitoring tool that would capture the essential elements of the TOP program as described in the TOC of the program and also matched the execution of regular TOP home visits. Using an iterative approach we employed concepts and made several adaptations and tested several options using video vignets from home visits. To distill a complex intervention into a practical number of measurable items and for the tool be readily available for later use we recommend this partnership between researchers, educators and practioners. Although we used an iterative process and did not determine the needed phases in advance, creating the fidelity tool ended in a 3-step process. Practitioners are very sensitive to the challenges of fellow interventionists and we would therefore recommend independent raters and extensive training for practioners. Forgatch also described this bias and suggested raters that are unfamiliar with the trainees (Forgatch, Patterson, & DeGarmo, 2005).

4.4. Limitations and future research directions

In this study, we addressed and contributed to the gap in fidelity measurement tools for interventions at scale. We developed the fidelity tool to measure two constructs; adherence with the delivery of the key strategies and competence in intervention delivery and added scores for parent availability and total impression. However, to measure the full construct of intervention fidelity and relate this to intervention outcomes, the other constructs such as dose (amount of intervention delivered) need to be included as well (An, Dusing, Harbourne & Sheridan, 2020). Since the actual delivery of the home visits is registered, this can easily be added. For measuring parental responsiveness, we used one overall score (Item 11) to describe the parental engagement with the interventionist and active participation in the interaction with their child. Since parental responsiveness is a target in our program and all of the key strategies and intervention specific competences are designed to let the parents be in the lead and strengthen parent-child interactions parental responsiveness can vary and improve during the trajectory and should be measured more than once.

A limitation that may have affected our reliability results is the restrictive 3-point Likert scale. Although this seemed to be the optimum number of response categories in the development of the tool, it caused lower variance in certain items while coding the videos. ICC values are likely to have been reduced due to the restriction of range in scores. Availability of more diverse videos with various distribution in scores could have resolved this issue.

4.5. Future research directions

The criteria to determine what an acceptable level of fidelity is, still needs to be determined. This is an important next step when systematically examining the associations between program fidelity and outcomes. The TOP program has been carefully implemented but may need adaptations in response to changes in participant needs, new insights, and other resources available. The FITT has great value to support further development of the intervention and enlarge its impact since it allows to monitor the pre-defined changes.

5. Conclusion

This study describes the iterative development of the FITT, an intervention-specific fidelity measure for the TOP program. The co-creative and iterative process led to a tool that objectifies what is done in the home visits, thereby adding transparency and transferability to the TOP program. The high association between the total impression item and subscales shows that the FITT identifies and captures the unique execution of the TOP intervention.

By working together with practitioners, researchers, and lecturers, the FITT became a relevant and reliable fidelity tool and will serve for future evaluation of the TOP program and contribute to measuring the professional development of the TOP interventionist and the continuous quality improvement efforts.

We would strongly recommend that when a new intervention is developed, a fidelity measure should be developed simultaneously to improve the study's internal validity and safeguard the quality of the intervention during the implementation process.

Ethical review

The Medical Ethical Review Committee (METC) of the Academic Medical Centre (AMC) Amsterdam has concluded that the current study was not covered by the Law Medisch Wetenschappelijk Onderzoek (WMO) and provided therefore a waiver.

Lessons learned

Developing a usable and reliable fidelity tool for a process-oriented intervention was challenging. We learned that development of the fidelity measure took time and effort, however the co-creative process offered many additional benefits. The open discussions between developers and educators and interventionists about detected deviations, meaningful adaptations, and difficulties in translating the key strategies into practice facilitated its uptake in the educational program and gave ideas for maintaining and improving the effectiveness of the program.

Creating the fidelity tool from the existing detailed and specific Theory of Change, training materials, and translating the theoretical base and key strategies into objective and quantifiable components helped define the intervention's target. Identifying the necessary specific competence skills to deliver an effective home visit was complex but enhanced our understanding of the strength of our program and interventionists. We encourage intervention developers to simultaneously develop a fidelity measure. This enables monitoring the actual program delivery by interventionists and helps to safeguard the quality of the execution.

Funding

Phase I of this study was supported by Harvard University, the Center on the Developing Child's R&D platform. Agreement number: 256529-5108674. Funding bodies played no role in the study's design, data interpretations, or manuscript writing.

CRediT authorship contribution statement

Monique Flierman, primary responsible for conceptualization, methodology, creation tool developing data collection and analysis and writing original draft of the manuscript. Eline Vriend; conceptualization, project administration, data curation, review and editing,. Raoul H. H. Engelbert, supervision, review and editing. Aleid Leemhuis, review and editing. Martine Jeukens-Visser, methodology, formal analysis, funding acquisition, review and editing.

Acknowledgements

The authors would like to thank Esther van der Heijden and Marjolein van Velsen, TOP lecturers, for their significant contribution, and all TOP therapists and families involved in the study. A special thanks to the team of Frontiers of Innovation of the Center on the Developing Child for facilitating this study.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.evalprogplan.2023.102299.

References

An, M., Dusing, S. C., Harbourne, R. T., & Sheridan, A. M. (2020). What really works in intervention? Using fidelity measures to support optimal outcomes. *Physical Therapy, 100*(5), 757–765. https://doi.org/10.1093/ptj/pzaa006

An, M., Nord, J., Koziol, N. A., Dusing, S. C., Kane, A. E., Lobo, M. A., & Harbourne, R. T. (2021). Developing a fidelity measure of early intervention programs for children with neuromotor disorders. *Developmental Medicine and Child Neurology, 63*(1), 97–103. https://doi.org/10.1111/dmcn.14702

Anderson, P. J., Treyvaud, K., & Spittle, A. J. (2020). Early developmental interventions for infants born very preterm - what works? *Semin Fetal Neonatal Med, 25*(3), Article 101119. https://doi.org/10.1016/j.siny.2020.101119

Asan, O., & Montague, E. (2014). Using video-based observation research methods in primary care health encounters to evaluate complex interactions. *Informatics in Primary Care, 21*(4), 161–170. https://doi.org/10.14236/jhi.v21i4.72

Askeland, L. E., Forgatch, M. S., Apeland, A., Reer, M., & Gronlie, A. A. (2019). Scaling up an empirically supported intervention with long-term outcomes: the nationwide implementation of generationPMTO in Norway. *Prevention Science, 20*(8), 1189–1199.

Bastick, E., Bot, S., Verhagen, S. J. W., Zarbock, G., Farrell, J.&, Brand-de Wilde, et al. (2018). The development and psychometric evaluation of the group schema therapy rating scale - revised. *Behavioural and Cognitive Psychotherapy, 46*(5), 601–618. https://doi.org/10.1017/S1352465817000741

Breitenstein, S. M., Fogg, L., Garvey, C., Hill, C., Resnick, B., & Gross, D. (2010). Measuring implementation fidelity in a community-based parenting intervention. *Nursing Research, 59*(3), 158–165. https://doi.org/10.1097/NNR.0b013e3181dbb2e2

Van Hus, J., Jeukens-Visser, M., Koldewijn, K., Holman, R., Kok, J. H., Nollet, F., & Van Wassenaer-Leemhuis, A. G. (2016). Early intervention leads to long-term developmental improvements in very preterm infants, especially infants with bronchopulmonary dysplasia. *Acta Paediatr, 105*(7), 773–781. https://doi.org/10.1111/apa.13387

Center on the Developing Child; IDEAS impact framework, 2015. Retrieved on 01–03-2023 from: ⟨https://tps://ideas.developingchild.harvard.edu/evaluation/⟩.

Cross, W., West, J., Wyman, P.A., Schmeelk-Cone, K., Xia, Y., Tu, X., Teisl, M.,et al. 2015) Observational measures of implementer fidelity for a school-based preventive intervention: development, reliability, and validity. *Prev Sci.16* (1)122–132 doi:10.1007/s11121–014-0488–9.

Cross, W. F., & West, J. C. (2011). Examining implementer fidelity: Conceptualizing and measuring adherence and competence. *Journal of Children's Services, 6*(1), 18–33. https://doi.org/10.5042/jcs.2011.0123

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: a review of research on the influence of implementation on program outcomes and the factors affecting implementation. *The American Journal of Community Psychology, 41*(3–4), 327–350.

Fixsen, D. L., Blase, K. A., Naoom, S. F., & Wallace, F. (2009). Core implementation components. *Research on Social Work Practice, 19*(5), 531–540.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33*(3), 613–619.

Forgatch, M. S., & DeGarmo, D. S. (2011). Sustaining fidelity following the nationwide PMTO implementation in Norway. *Prevention Science, 12*(3), 235–246. https://doi.org/10.1007/s11121-011-0225-6

Forgatch, M. S., Patterson, G. R., & DeGarmo, D. S. (2005). Evaluating fidelity: predictive validity for a measure of competent adherence to the Oregon mode l of parent management training. *Behavior Therapy, 36*(1), 3–13.

Gearing, R. A., El-Bassel, N., Ghesquiere, A., Baldwin, S.,J., Gillies, J., & Ngeow, E. (2011). Major ingredients of fidelity: a review and scientific guide to improving quality of intervention research implementation. *Clinical Psychology Review, 31*(1), 79–88. https://doi.org/10.1016/j.cpr.2010.09.007

Goering, P., Veldhuizen, S., Nelson, G. B., Stefancic, A., Tsemberis, S., Adair, C. E., et al. (2016). Further validation of the pathways housing first fidelity scale. *Psychiatric Services, 67*(1), 111–114. https://doi.org/10.1176/appi.ps.201400359

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1*(1), 77–89.

Huth-Bocks, A. C., Jester, J. M., Stacks, A.M.,Muzik, M. K., Rosenblum.L, & Michigan, R. (2020). Infant mental health home visiting therapists' fidelity to the Michigan IMH-HV model in community practice settings. *The Infant Mental Health Journal, 41*(2), 206–219.

Jeukens-Visser, M., Koldewijn, K., van Wassenaer-Leemhuis, Flierman, M., Nollet, F., & Wolf, M. J. (2020). Development and nationwide implementation of a postdischarge responsive parenting intervention program for very preterm born children: The TOP program. *Infant Ment Health J*. https://doi.org/10.1002/imhj.21902

Koldewijn, K., van Wassenaer, A., Wolf, M. J., Meijssen, D., Houtzager, B., Beelen, A., … Nollet, F. (2010). A neurobehavioral intervention and assessment program in very low birth weight infants: outcome at 24 months. *J Pediatr, 156*(3), 359–365. https://doi.org/10.1016/j.jpeds.2009.09.009

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *The Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Martin, M., Steele, B., Lachman, J. M., & Gardner, F. (2021). Measures of facilitator competent adherence used in parenting programs and their psychometric properties: a systematic review. *Clinical Child and Family Psychology Review, 24*(4), 834–853. https://doi.org/10.1007/s10567-021-00350-8

Mowbray, C. (2003). Fidelity criteria: development, measurement, and validation. *The American Journal of Evaluation, 24*(3), 315–340. https://doi.org/10.1016/s1098-2140(03)00057-2

Parham, L. D., Roley, S. S., May-Benson, T. A., Koomar, J., Brett-Green, B., Burke, J. P., et al. (2011). Development of a fidelity measure for research on the effectiveness of the Ayres Sensory Integration intervention. *American Journal of, 65*(2), 133–142. https://doi.org/10.5014/ajot.2011.000745

Perined. (2020). Perinatale zorg in Nederland anno 2019. Retrieved from ⟨https://assets.perined.nl/docs/aeb10614–08b4–4a1c-9045–8af8a2df5c16.pdf⟩.

Di Rezze, B., Law, M., Eva, K., Pollock, N., & Gorter, J. W. (2013). Development of a generic fidelity measure for rehabilitation intervention research for children with physical disabilities. *Developmental Medicine and Child Neurology, 55*(8), 737–744. https://doi.org/10.1111/dmcn.12114

Schindler, H. S., Fisher, P. A., & Shonkoff, J. P. (2017). From innovation to impact at scale: Lessons learned from a cluster of research-community partnerships. *Child Development, 88*(5), 1435–1446. https://doi.org/10.1111/cdev.12904

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420.

Song, M. K., Happ, M. B., & Sandelowski, M. (2010). Development of a tool to assess fidelity to a psycho-educational intervention. *The Journal of Advanced Nursing, 66*(3), 673–682. https://doi.org/10.1111/j.1365-2648.2009.05216.x

Tidmarsh, G., Whiting, R., Thompson,J.L, J., & Cumming, J. (2022). Assessing the fidelity of delivery style of a mental skills training programme for young people experiencing homelessness. *Evaluation and Program Planning, 94*, Article 102150. https://doi.org/10.1016/j.evalprogplan.2022.102150

Toomey, E., Matthews, J.,S., Guerin, S., & Hurley, D. A. (2016). Development of a feasible implementation fidelity protocol within a complex physical therapy-led self-management intervention. *Physical Therapy, 96*(8), 1287–1298. https://doi.org/10.2522/ptj.20150446

Monique Flierman is an educator and researcher in the Department of Rehabilitation at the Amsterdam University Medical Center. She is a pediatric physical therapist and has her master's degree in Education Sciences. She serves as co-director of the Expertise Centre for Premature Infants (EOP-nl) at the Amsterdam UMC. Her expertise and research focuses on the implementation of early intervention programs for premature born infants.

Eline Vriend is a junior researcher in the Department of Rehabilitation at the Amsterdam University Medical Center. She has a bachelor's degree in health sciences and a master's degree in pedagogical sciences from the VU University Amsterdam. She is involved in research projects on early intervention for preterm infants and their parents, and in the continuous evaluation of the TOP program.

Aleid Leemhuis is pediatrician and head of Neonatal Follow up at Amsterdam UMC and member of the EOP in a consultant role. Her research focuses on primary, secondary and tertiary prevention of adverse outcomes in high risk neonates. As such she was involved with the development and research related to the TOP intervention right from the start.

Raoul Engelbert is Professor in Pediatric Physical Therapy at the Faculty of Health of the Amsterdam University of Applied Science, the Department of Rehabilitation Science and Pediatrics of the Amsterdam University Hospital AMC in Amsterdam. Special interest in chronic diseases in childhood (functional ability, physical fitness, performance and capacity as well as participation in society and leisure activities; diagnostics and interventions). Special interest groups: neonatology, cardiopulmonology, connective tissue disorders in childhood and adolescence. Together with a research team they guide 17 PhD candidates. He chairs the scientific committee of the Dutch Association of Pediatric Physical Therapy.

Martine Jeukens-Visser is a senior researcher in the Department of Rehabilitation at the Amsterdam University Medical Center. Her research focuses on the development of very preterm born infants and early intervention for preterm infants and their parents. She is involved in the implementation and continuous evaluation of the TOP program.