emerald PUBLISHING

Journal of Children's Se

The Reflective Fostering Programme Fidelity Rating Scale: Development and Inter-Rater Reliability

Journal:	*Journal of Children's Services*
Manuscript ID	JCS-01-2022-0002.R1
Manuscript Type:	Research Paper
Keywords:	programme fidelity, foster care, treatment fidelity assessment, Reflective Fostering Programme, parenting programme, adherence scales

The Reflective Fostering Programme Fidelity Rating Scale: Development and Inter-Rater Reliability

Abstract

An accurate assessment of programme fidelity is critical to the reliability, validity and scale-up of the results of any intervention research study. **Purpose:** This study describes the development of the 14-item Reflective Fostering Fidelity Rating (RFFR), an observational rating system to evaluate model fidelity of group facilitators in the Reflective Fostering Programme (RFP), a mentalization-based psychoeducation programme to support foster carers. We assess usability, dimensionality, inter-rater reliability, and discriminative ability of the RFFR. **Methodology:** Eighty video clip extracts documenting 20 RFP sessions were independently rated by four raters using the RFFR. The dimensionality of the RFFR was assessed using principal components analysis. Inter-rater agreement was assessed using the intraclass correlation coefficient. **Findings:** The proportion of missing ratings was low at 2.8%. A single principal component summarised over 90% of the variation in ratings for each rater. The inter-rater reliability of individual item ratings was poor to moderate, but a summary score had acceptable inter-rater reliability. We present evidence that the RFFR can distinguish RFP sessions that differ in treatment fidelity. **Originality**: This is the first investigation and report of the RFFR's validity in assessing the programme fidelity of the RFP. The paper concludes that the RFFR is an appropriate rating measure for treatment fidelity of the RFP and useful for the purposes of both quality control and supervision.

Keywords: programme fidelity, foster care, treatment fidelity assessment

Article Classification: Research Paper

Introduction

It is estimated that more than 1.4 million children are in formal care (including foster or kinship care, and residential care) around the globe (Petrowski *et al.*, 2017), a population who typically experience disproportionately higher rates of physical, mental and developmental health problems alongside healthcare needs that are unmet (Lindley and Slayter, 2019). Children in care are "one of the most vulnerable and disadvantaged groups in our society" (NICE, 2010, p.10), with over 65% of children in England having experienced abuse or neglect prior to placement (NICE, 2021).

Systematic reviews have indicated that interventions for foster carers (including kinship or connected carers) such as Treatment Foster Care may play an important part in encouraging positive outcomes for the children in their care (e.g., Turner and Macdonald, 2011). A number of interventions aim to mitigate negative outcomes for children in care by improving parenting practices, and thereby promoting the child's emotional and behavioural well-being (Vanschoonlandt *et al.*, 2012). To date most parenting programmes developed for foster carers have taken a primarily behavioural approach (sometimes including an attachment perspective), and although there is some evidence of effectiveness, trials have been inconclusive (Bergstrom *et al*., 2020), especially with regard to longer-term outcomes and impact on key areas, such as placement stability.

The Reflective Fostering Programme (RFP, Redfern *et al*., 2018) is a mentalization-based psychoeducational parenting programme for foster carers of children aged four to thirteen. Unlike many parenting programmes, Reflective Fostering focuses on supporting and improving the carer-child relationship, rather than teaching specific parenting strategies, with an emphasis on the foster carer's own responses, thoughts and feelings. The programme consists of 10, three-hour group sessions which are delivered weekly by two facilitators, one

of whom is a social care staff worker, and the other a foster carer. Both of these facilitators receive a three-day training to deliver the programme, alongside a weekly consultation session to support programme fidelity. Each session follows specified content designed to provide carers with practical tools to increase their capacity for mentalizing and reflective parenting. Mentalizing refers to the ability of carers to understand their own, and their child's, mental states and how these may underlie behaviours (Slade, 2005). By attending to their own state of mind and experiences in this way, carers are able to better manage their own feelings and respond to the needs and behaviours of the child in their care in an effective manner. A fuller description of the RFP can be found in Redfern *et al*. (2018).

Preliminary evaluations of the programme show that it is effective and well-received by participating foster carers. A feasibility study of the programme confirmed that its delivery and the associated training of facilitators was feasible and preliminary evaluations produced positive outcomes (Midgley *et al*., 2019). Results from this study showed a significant reduction in parenting stress (Parenting Stress Index, Short Form; Abidin, 1995) and child difficulties (Strengths and Difficulties Questionnaire; Muris *et al.*, 2003) following the completion of the programme. A further pilot evaluation of the programme has since been completed with comparable outcomes to the initial feasibility study (Midgley *et al*., 2021a) and a large-scale randomized control trial of the programme is currently underway in the U.K. (Midgley *et al*., 2021b).

Programme Fidelity

Fidelity of an intervention can broadly be defined as "the intervention being delivered as intended by the program developers and consistent with the program model" (Breitenstein *et al*., 2010, p. 2). Some authors distinguish between adherence (the degree to which prescribed intervention activities and strategies are used) and competence (the skill and style with which

the program is delivered) (Martin *et al.*, 2021). Assessing fidelity is important not only to support the training and development of practitioners (called 'facilitators' in the RFP), but also as a way of assessing whether the intervention has been delivered in a way that is consistent with the core principles and techniques of the approach. Without being able to assess this, one cannot be sure whether any identified outcomes can be attributed to the intervention (Moore *et al*., 2015; Oakley *et al.*, 2006).

Examining intervention fidelity is also important because the use of study protocols and intervention manuals alone cannot guarantee that a programme will be delivered in the way it is intended (Ogrodniczuk and Piper, 1999). For example, "drift" from the programme specification or manual may occur when a programme is delivered by new facilitators (Bywater *et al.*, 2019), or over time. It is therefore common practice to use fidelity criteria to monitor fidelity of an intervention (Mowbray *et al*., 2003). Routine monitoring of programme delivery is also of great importance for ensuring consistency, particularly in multi-site studies (Paulson *et al.,* 2002). Additionally, fidelity measures can also serve as records of programme delivery to refer to, for example, when investigating why an intervention is ineffective (Chen, and Chen, 1990; Hohmann and Shear, 2002, Mowbray *et al*., 2003). Some previous research has demonstrated a positive association between assessed fidelity and programme outcomes, with greater fidelity predicting improved outcomes for participants (e.g. Durlak and DuPre, 2008; Eames *et al*., 2009; Elliott and Mihalic, 2004; Kam *et al.*, 2003). Fidelity measures have also been shown to predict outcomes during replications of model delivery (Paulson *et al*., 2002). When key components are absent, results have been less positive (Bond *et al.*, 2000), demonstrating the usefulness of fidelity criteria in measuring true effectiveness of interventions. Furthermore, information regarding programme fidelity is useful to facilitate further development of interventions. Not only can fidelity ratings allow researchers to identify barriers to delivery (e.g. de Vet *et al*., 2015), but they can also reveal how the programme alters

over time and in different settings (Goulet *et al.*, 2018). This is important to consider when planning for programme delivery into wider contexts and across different populations, as well as in programme reports for funders.

Approaches to Assessing Programme Fidelity

It is reasonable to assume that the most effective way of achieving fidelity in programme delivery is by ensuring facilitators or programme leaders receive training and ongoing support and consultation throughout programme delivery. Such consultation and support can be used to monitor programme fidelity through several different methods, particularly using observational and self-report measures (Breitenstein *et al*., 2010; Lorencatto *et al.*, 2013; Toomey *et al.*, 2017; Walton *et al.*, 2017).

Self-report measures of programme fidelity often include session checklists; a report completed by the session leader detailing if specific content has been delivered, what materials were used, and if specified tasks were set. Such measures are easy to complete routinely, but the information collected is often limited and open to subjective bias (Breitenstein *et al*., 2010; Bywater *et al*., 2019). Additionally, such reports do not always correlate with reports provided from independent supervisors or raters (Eames *et al*., 2008). Because of these limitations, observational techniques (i.e., video/audio recordings of sessions to be reviewed at a later time) are often viewed as the gold standard to assess fidelity (Borrelli, 2011; Lorencatto *et al*., 2013; Walton *et al*., 2017).

Fidelity Assessment in Parenting Programmes

A recent systematic review of fidelity measures in parenting programmes (Martin *et al*., 2021) identified 65 such measures, none of which were specifically identified as exclusively targeted at assessing parenting programmes for foster carers. However, some of the identified programmes, such as the Triple P, Incredible Years and Parent Management Training

Programme – Oregon Model interventions have been used with foster carer populations and evaluated for fidelity (Maaskant, van Rooij, Overbeek, Oort, & Hermanns, 2016; Job *et al.,* 2022; McDaniel, Braiden, Onyekwelu, Murphy & Regan, 2011). For 30 of these measures, data regarding psychometric properties have been reported (Martin *et al*., 2021). This review found that there is significant variation both in the nature of fidelity measures utilised in parenting programmes and also the extent to which the psychometric properties, validity and reliability of these measures have been established. In terms of the measures themselves, of the studies that reported on the data used to assess fidelity, there was a relatively even split between those that utilised observational data and those that used non-observational methods such as rating from memory. There was also diversity in the individuals that conducted fidelity assessments, including session facilitators, researchers, third parties, attending parents, supervisors or a combination of the above. This review also found that while some studies reported on indicators of the reliability and validity of these measures others failed to include this information. To assess the quality of these studies the authors developed a Study Risk of Bias and Quality Checklist developed from the COSMIN guidelines (Martin *et al*., 2021; Mokkink *et al*., 2010 a/b, Terwee *et al*., 2007). This checklist assesses the presence of criteria that determine the quality of determined measures for example, random selection of rating material, use of independent assessors and consideration of reactivity. A Measure Practicality Checklist was utilised by Martin et al. (2021) to assess measure sustainability, utility and availability, including an assessment of how much training was needed to make use of each fidelity measure. Furthermore, the review examined whether studies reporting on fidelity measures in parenting programmes reported on measures of internal, test-re-test and inter-rater reliability, as well as reporting on construct and convergent validity. One typical shortcoming of papers reviewed were the use of non-observational measures, leading to lower reliability and introduction of bias due to self-assessment by raters. Furthermore, while several studies

reported on the inter-rater reliability of measures of fidelity, few assessed other forms of reliability or validity. It is important to note that several studies evaluating interventions specifically targeted at foster carers, such as the Fostering Changes and Fostering Connections Programmes have highlighted the need for greater accuracy in measuring fidelity for interventions with this population (Moody *et al.,* 2020; Lotty, Bantry-White & Dunn-Galvin, 2020).

The Current Study

This paper introduces the Reflective Fostering Fidelity Rating (RFFR) system, developed as a measure of programme fidelity during the pilot evaluation phase of the RFP.

The primary aims of the present study were (i) to test the capacity to rate the programme sessions based on the developed scheme, and (ii) to investigate inter-rater agreement and the degree of consistency of ratings across dimensions of programme fidelity. Our research questions were as follows:

(1) Usability: Can raters use the 14 RFFR items to rate fidelity to the RFP?

(2) Inter-rater agreement: How reliable are ratings of facilitator fidelity to the RFP across different raters?

(3) Dimensionality and internal consistency: Are fidelity ratings pertaining to different dimensions (aspects) of the programme sufficiently consistent with one another to justify summarising the ratings in a single measure of fidelity, or is it more appropriate to consider the concept of fidelity to the programme as multidimensional?

(4) Discriminative ability: Can RFFR ratings distinguish between sessions that differ in the extent of fidelity? This refers to the capacity of the ratings to distinguish sessions that are 'on model' from those that are 'off model' (i.e., sessions which are delivered according to the intervention manual, and those which are not).

Method

Setting for the study

The data used in this study were collected as part of a pilot evaluation of the Reflective Fostering Study (for full details, see Midgley *et al*., 2021a). The study was approved by UCL Ethics Committee (Approval ID Number: 14653/001).

The Reflective Fostering Fidelity Rating (RFFR) System

The development of the scale (item selection, item definitions, quality descriptors, rating procedures, etc.) was performed in a dialogue between the programme developers and the research team involved in the feasibility evaluation of the programme. This involved reviewing the programme manual and extracting the key components that were considered essential to delivering the programme 'on model' and identifying clear enough descriptions of behavioural markers for these components. The first version of the scale was based on this review of the Reflective Fostering Manual and consisted of 21 items rated on a 4-point response scale. It aimed to identify the key elements of the programme (i.e. mentalization, the promotion of reflective capacity, and supporting the enhanced monitoring of one's 'emotional temperature'), organised in relation to some of the key elements of the programme, such as the 'Professional APP', the 'Carer APP' (APP referring to the mentalizing stance of attention and curiosity (A), perspective taking (P) and providing empathy (P)) and the 'Carer Map', which is core tool of the Reflective Parenting Model, designed to help carers identify their current state of mind, what is influencing this and the impact of past family history and early experiences on their caregiving (see Redfern *et al.,* 2018).

To test face validity and usability, this initial version of the scale was rated by clinical consultants who had been involved with the development of the programme, using entire video

recordings of three-hour sessions from the initial feasibility study. Three consultants involved with the development of the programme rated each session offered to four groups, then met to discuss any issues that arose from the rating process. At the end of the pilot study, some changes were made to both the manual and the Fidelity scale, in order to ensure that the items were comprehensive and could be rated based on observation of a session. This primarily involved finding ways to identify behavioural markers of key elements of the programme, e.g. 'noticing and naming changes in emotional temperature' became a more explicit way to describe the key role of supporting emotion regulation.

Since the original 21 items were chosen by the developers of the programme, the items had a high degree of content validity, covering a wide range of interventions. However, the first use of the RFFR revealed a need for some clarifications and redefinitions of the scale items and their descriptors. Some items were felt to be redundant, while others needed to be simplified or their description be revised based on changes made to the Manual. For example, where in the initial 21-item system there were several items relating to facilitators covering of key session activities and themes, these were found to be consistently rated with the same score. Therefore, these items were consolidates into Item 1 'consistent and explicit focus on the aims / themes of the session'.

The scale revision was made in collaboration between the clinical and research team working on the programme to ensure that the items were able to capture the core elements of the intervention, while at the same time being formulated in a language which was not overly theoretical or abstract. This work resulted in the selection of 14 items. The response scale was revised to a five point scale, from 1=poor, to 5=excellent, with 3 indicating an 'adequate' level of programme fidelity. By using a scale based on the quality of delivery, each item is intended to assess 'competent adherence', rather than rating separately for competence and adherence,

as some other fidelity measures have done. This was based on our experience that, in the context of this programme, it was of limited value to try and assess adherence and competence separately, as a particular activity (such as paying attention and showing curiosity) could not be considered 'adherent' if it was not also done competently. The broader concept of 'fidelity' or 'competent adherence' was considered to be more appropriate, where each item is rated on a single scale, with regard to whether it was delivered as intended and in a way that was consistent with the programme model. As part of the work on the coding manual, descriptions and examples to support the ratings of each item were developed. The items from this updated, final version of the scale are displayed in Table I. See Appendix 1 for the comprehensive coding manual.

[Table I near here]

Although initial development work focused on ratings of whole sessions, previous studies have demonstrated that fidelity can be assessed based on relatively short extracts of sessions (e.g. Weck *et al*., 2011) and that this increases the chances of fidelity measures being used in practice (Weck *et al*., 2014). The updated version of the RFFR was, therefore, designed to be rated on the basis of a review of 20 minutes of video, made up of four purposively sampled 5-minute video clips from each session. Group facilitators are asked, at the end of each session, to select three clips, each 5 minutes in length. They are told that these clips should, respectively, include: a) the introduction of the session aims and opening Mind Check activity (a specific reoccurring feature of each session of the programme); b) a segment where they felt that things were 'going well'; and c) a challenging segment where it was felt the group was going less well. A fourth clip is then selected by the person doing the rating, in which there is an example of the facilitators leading one of the activities specified for that session that has not been captured in the other selected clips. The clips are purposefully chosen to ensure that the above

mentioned activities/events are included. The research team then rated a selection of sessions that had previously been rated based on review of the whole session, this time doing the ratings based on these four clips, and compared these ratings to the earlier ratings of whole sessions. Session sampling in this way was utilised in order to reduce the likelihood of assessor bias within the subsequent ratings (Martin *et al*., 2021; Mokkink *et al.,* 2010a; Mokkink *et al*., 2010b; Terwee *et al*., 2007). This exercise indicated that the RFFR could be rated on the basis of the four clips, with only a small number of adaptations to the coding manual. This version of the RFFR is the one that is evaluated in the current study. The measure was designed with the ultimate goal of it being used by consultants to Reflective Fostering programme session facilitators as part of the supervision process.

Data Collection

Video recordings of programme sessions were taken as part of the second pilot of the RFP (see Midgley *et al*., 2021). The video recordings were made on tablets by facilitators during all programme sessions, except where technological issues prevented this. Recordings came from five different groups, each led by a different pair of facilitators. Each video recording was approximately three hours long, the length of the entire session, and the camera was focused on the two facilitators rather than the group as a whole, to provide some anonymity for participants. The sessions were purposively selected to reflect a mix of the different groups (group A = 5, group B = 3, group C = 2, group D = 5, group E = 5) and sessions (sessions 1-5 = 11, sessions 6-10 = 9) where possible given recording availability. Four clips were taken from each of the 20 sessions. The selection of clips is described in the previous section.

Ratings were provided by four independent raters, who were all clinicians and/or research psychologists involved in the development and/or training and supervision of the RFP. A half-

day training was organised on the RFFR, in which items were reviewed and four 5-minute clips were watched, rated independently and then discussed. Each rater rated all twenty sets of clips, and was blinded to the selection process of the clips.

Statistical Analysis

We analysed ratings of the 20 clip sets, each rated independently by four raters, using all 14 items on the RFFR. This gave us 1,120 data points. All analyses were conducted using the R software for statistical computing (R Core Team, 2019), v. 3.6.0. The following R packages were used for data processing and analysis: tidyverse, psych, MBESS, and lme4.

(1) We examined the usability of the RFFR system by documenting summary statistics on each of the 14 items for each of the 4 raters separately, including means, standard deviations, and number and proportion of missing ratings (i.e. where the rater decided they could not rate an item on a particular selection of clips).

(2) We assessed the dimensionality of the 14 items using multilevel exploratory factor analysis (Reise et al 2005). This method takes account of the multilevel structure of our data: ratings are clustered within raters. In multilevel factor analysis, the total covariance matrix is decomposed into two sources: the within-cluster covariances and the between-cluster covariances. In our case, we performed a factor analysis of the pooled within-rater covariance matrix. We did not analyse the between-rater covariance matrix, since this does not address our research question (Reise et al 2005). To ensure robustness of our conclusions to different assumptions of missing value generating mechanisms, we performed this analysis in two ways: (i) using clip sets with complete ratings only (complete cases analysis, n = 50), (ii) using all clip sets, imputing missing ratings as the mean of the valid ratings for that rater and that clip set (n = 80). Dimensionality was assessed by considering the eigenvalues of the covariance matrix,

including via a scree plot, and the total variance summarized by sensible factor models. We assessed the reliability of the resulting dimension by calculating McDonald's ω (McDonald, 1999) for the pooled within-rater covariance matrix, using the MBESS package in R (Kenney, 2022).

(3) Inter-rater agreement was assessed using the intraclass correlation coefficient for single judges, absolute agreement, and fixed effects for raters. This corresponds to the mixed effects model ICC(3,1) in Shrout and Fleiss (1979) – see also McGraw and Wong (1996). We calculated this ICC separately for each item. Since the analysis under (2) suggested that RFFR ratings can be summarized using a single dimension (see Results section), we also assessed the inter-rater agreement on an overall RFFR score calculated as the sum of the 14 item ratings. The "ICC" function from the psych package (Revelle, 2021) was used to calculate the coefficients and their confidence intervals. The mixed effects models used in the calculation can accommodate missing values. The results are valid under the assumption that ratings are missing at random given the observed ratings. We also calculated percentage agreement for each item, as well as overall percentage agreement. In the context of four raters, percentage agreement is calculated as the ratio of actual agreements between any two raters over the number of possible agreements (that is, the sum of pairwise comparisons). Since we report on missing ratings separately in the descriptive analysis, missing ratings were not considered in the percentage agreement calculation (that is, a missing rating was not counted as a disagreement with another rater who did give a rating).

(4) Finally, we assessed the ability of the RFFR summary score to distinguish between sessions with higher or lower fidelity, by tabulating the scores for the 20 clip sets, separately for each rater as well as averaging over the four raters.

The full data set of ratings, and R code to reproduce all analyses, are deposited in the UCL Research Data Depository.

Results

Description and Rateability

Table II shows descriptive statistics for the 14 items, separately for each rater and overall. Of the 1,120 expected ratings, 32 were missing (2.8 %). Most of these related to item 2 (use of Mind Checks), which had 27 missing ratings (33.8 % of the 80 expected ratings). All other items had either 0 or 1 missing rating. Of the raters, Rater A had the most missing values (16 out of 280, or 5.7 %). The other raters had 5 or 6 missing ratings each, all but one of which related to item 2. Overall, then, the proportion of missing values was small for all items except for item 2, and there may have been some difference between Rater A and the remaining raters regarding the perceived rateability of this item.

Table II provides preliminary indication that raters varied somewhat with respect to both central tendency and scale (variance) of their ratings. Considering the summary score, within-rater means varied from 36.95 (rater D) to 40.62 (rater A). Standard deviations varied from 7.61 to 12.06.

[Tables II & III near here]***Inter-item Correlations and Dimensionality***

Table III reports the pooled within-rater Pearson correlations of the 14 items (using mean replacement – the results using complete cases are very similar). All correlations are positive, as expected. The smallest observed correlation is r = 0.05 (items 1 and 10), the largest is r =

0.86 (items 8 and 14). The average of the 91 correlations is 0.47 if using mean replacement for missing item values (0.46 for the complete cases).

[Figure 1 near here]

A scree plot of the Eigenvalues of the pooled within-rater covariance matrix is shown in Figure 1. The first eigenvalue is by far the largest (7.2 and 7.1 respectively using data with mean replacement for missing values or complete cases). The second eigenvalue is 1.6 (in both analyses). All other eigenvalues are smaller than 1. The screeplot suggests that either a one-factor or a two-factor model may be sensible. We therefore inspected both models. Promax rotation was used to allow for correlated factors in the two-factor model.

[Table IV near here]

Table IV shows the estimated loadings for the one- and two-factor solutions, using data with mean replacement for missing values to calculate the covariance matrix. The results are essentially identical if complete cases are used instead to calculate the covariances. In the one-factor model, all items load highly on the single factor, which accounts for 48.6 % of the total variance. The two factor model features two highly correlated factors (r = 0.557), accounts for 51.8 % of the total variance, and has almost simple structure: Items 1-7, 13 and 14 load highly on Factor 1 only, items 9-12 on Factor 2 only; but item 8 loads (moderately) highly on both factors. Factor 1 may thus largely relate to the fidelity of session content and of group process facilitation (the 'what' of the session), while Factor 2 may capture the facilitator mentalizing stance and the encouragement of such a stance in participants (the 'how'). However, the model had no simple structure, and it is unclear how to interpret the role of item 8 in this regard.

The results of the factor analysis are somewhat ambiguous: there is some evidence for the existence of two factors, but the one-factor model is simpler and accounts for the data almost as well, since two-factor model explains only 3.2 percentage points more and appears to feature

at least one item that measures both factors. On balance, evidence in favour of the one-factor model is sufficiently strong to justify summarizing the RFFR system by a single summary score.

The reliability of the single summary score was estimated as McDonald's ω = 0.93 (95 % confidence interval: 0.91; 0.95), indicating excellent reliability.

Inter-Rater Agreement

Table V shows the observed intraclass correlation coefficients (and their 95% confidence intervals) for all items, as well as for the summary score. The point estimates for the 14 items vary from 0.28 (item 7) to 0.71 (item 14), suggesting that the reliability of individual item ratings is poor or moderate for all items except item 14. However, the summary score taken across the 14 items had a higher interrater reliability estimate of 0.74. ICCs above 0.7 are generally judged to indicate acceptable reliability. Percentage agreement was between 33 % and 48 % for items 1-13, but was considerably higher for item 14, at 57 %.

[Table V near here]

Ability to Discriminate Between On-Model and Off-Model Sessions

Table VI shows the RFFR summary scores for the 20 clip sets, by rater and averaged across raters. Mean replacement was employed for missing item ratings.

[Table VI near here]

There was considerable variation between clip sets in the rated fidelity. The lowest average summary score for a clip set was 26.1, which corresponds to an average item rating of 1.9. Recall that an item rating of 2 indicates 'inadequate' fidelity. The highest average summary score was 51.7, which corresponds to an average item rating of 3.7. An item rating of 3 indicates 'adequate', a rating of 4 indicates 'good' fidelity. On balance, there was thus

sufficient variation in the summary scores to be confident that the RFFR scale can distinguish RFP sessions that differ in treatment fidelity. Note that these ratings should not be taken as indications of the overall quality of the sessions in the pilot programme, since the sessions were purposely selected to represent a range of fidelity.

Discussion

The aims of the present study were to (i) test the capacity to rate clips from the Reflective Fostering Programme sessions using the Reflective Fostering Fidelity Rating (RFFR), and (ii) investigate inter-rater agreement and the degree of consistency of ratings across dimensions of programme fidelity.

Analysis of the data from four independent raters of 20 sets of clips from a group of purposively sampled sessions suggests that all items could be rated based on the clips provided. There was some indication that item 2, related to the use of a particular activity at the start of each session (the 'mind check'), was sometimes difficult to rate if the selected clips did not provide any evidence of whether this activity had been carried out. For this item, the coding manual did provide instructions on how to rate in this circumstance, but it appears that this had not been clear for all of those involved with the ratings. This finding emphasises the need for thorough and consistent training of raters using fidelity assessment tools to complement the development of thorough rating manuals for such measures.

Our analyses investigating the dimensionality of the 14 item-RFFR yielded somewhat ambiguous results. About half of the overall variation can be explained by a single factor, but there was some evidence for a two-factor model with highly correlated factors. These models deserve investigation with a new data set. Work is currently under way to collect a larger sample of RFFR ratings in the context of a randomized controlled trial (Midgley et al., 2021b),

and it is our intention to revisit the dimensionality question in that study. Pragmatically adopting a single factor for now, we found that a summary score calculated from all items with equal weights yielded a highly reliable measurement (McDonald's omega = 0.93).

There is a mixed picture with regard to inter-rater agreement. At an item level, inter-rater reliability was poor to moderate; but when a summary score was considered, inter-rater reliability was 0.74, which indicates an acceptable level. When put together with the evidence that there is considerable variation in the summary scores when comparing clips from 20 different sessions, this indicates that a summary score can be used as a reliable rating of whether a session is being delivered 'on model' or not; but that caution should be taken when examining at the level of individual items.

Comparisons with fidelity measures developed for use with other parenting programmes indicates some of the challenges in this field. For example, Martin *et al.*'s (2021) systematic review indicated that only 51.2% of the studies reporting on inter-rater reliability for parenting programme fidelity scales met the given reliability criteria (ICCs and Kappa's above 0.70 or Pearson's correlations above 0.80). It might be expected that inter-rater reliability would be even more challenging in the context of a mentalization-based parenting programme, as the training does not focus primarily on the use of specific techniques (e.g. teaching of a particular strategy for managing challenging behaviour), but rather on certain interpersonal skills, such as the ability to notice and name positive mentalizing between participants in the groups. However fidelity scales developed in the context of mentalization-based individual therapies (e.g. Karterud et al., 2013) have demonstrated fairly high levels of IRR, suggesting that the challenge may be more related to assessing fidelity of group-based parenting programmes, which are highly complex interventions involving multiple participants and group interactions, often facilitated by more than one professional. It is, therefore, crucial to develop fidelity measures that allow for flexibility in programme delivery and agility of parenting programmes

in different contexts while still reliably capturing competent adherence of facilitators or programme deliverers.

One of the obstacles identified by Martin et al. (2021) in the wider use of fidelity assessments is the fact that they are often time-consuming to complete. This has often meant that fidelity measures have been used in clinical trials, where funding to support this work may be available, but are under-utilised in routine practice. By developing a measure that can be rated on a selection of clips totalling 20 minutes (with an additional 5-10 minutes to rate the RFFP items), and demonstrating how the fidelity measure can be integrated into supervision of programme delivery, we hope that the RFFR has value beyond a research setting. However, we recognise that reliable use of the measure still depends on training in the RFFP (which in our experience takes about half a day, plus the opportunity to get feedback on practice ratings), as well as access to video-recordings of the Programme.

Strengths and Weaknesses of The Study

The development and validation of the RFFR is an important step towards evaluating the RFP. If proved to be an effective intervention in a randomised clinical trial currently underway (Midgley *et al*., 2021), it is anticipated that the RFP will be established as a professional qualification in which social care workers and foster carers can become accredited Facilitators. Therefore, it is vital to ensure that the programme is being delivered in a consistent and standardised way which reflects the aims and key components of the intervention. Development of a programme fidelity measure that is psychometrically sound and reliable is therefore necessary in evaluating the effectiveness of any new parenting programme.

Drawing on the Risk of Bias and Quality Checklist developed from COSMIN guidelines, several strengths of this study were identified. As mentioned earlier, four independent raters were included in the study which allowed estimation of inter-rater reliability with relatively

good precision (Hallgren, 2012; Walton *et al*., 2017). Similarly, the session sampling method utilised aimed to reduce the likelihood of assessor bias (Martin *et al*., 2021; Mokkink *et al.,* 2010a; Mokkink *et al*., 2010b; Terwee *et al*., 2007). In this study, sessions to be rated were blind selected to be a representative sample of all available recordings, with recordings rated for all different facilitated groups and from each session number. This fits with Martin and colleagues' (2010 a/b) recommendation for sessions chosen to be rated selected using a method that reduces selection bias (Ellenberg, 1994; Walton *et al*., 2017).

Certain limitations also need to be recognised. Although training on the measure was provided to raters, there is not yet a formal process for being accredited as a rater on the RFFR measure. However, it is notable that the RFFR has been shown in this study to be rateable by both clinicians and independent researchers. This suggests that the RFFR could feasibly also be completed by clinicians acting as supervisors or consultants to facilitators during group sessions, therefore enabling the RFFR to be integrated into the delivery of the RFP. It is anticipated that moving forward, consultant clinicians will use the RFFR to rate clips shared with them by facilitators and that these ratings will be used both for research purposes, as a measure of programme fidelity, and as a guide for supervision sessions. This is a strength of the RFFR as it increases the feasibility and sustainability of its use. It is notable that of the 22 studies examining measures of fidelity or adherence in parenting programmes by Martin and colleagues (2021) only 4 were rated by supervisors, making this a potentially under-utilised approach in this area, especially given its practical benefits.

A further limitation of this study is that there was not explicit consideration of facilitator reactivity in the collection of session recordings, where facilitators may behave differently because of data on their competent adherence being collected (Gardner, 2000; Kazdin, 1982). It is notable that all group sessions were recorded, not only those rated for this study, which allowed facilitators to habituate to the recording being taken and potentially reduce reactivity

to this in their behaviour (Martin *et al*., 2021; Kazdin, 1982). However, it is likely that facilitator reactivity may have influenced clip selection which may introduce inherent bias into the use of this measure. It is important to acknowledge that despite clips being selected primarily by facilitators for the RFFR, this study highlights that a range of the fidelity scale was used with some sessions rated as having inadequate fidelity and others rated as good or close to very good fidelity.

Conclusion

It is recognised that the mental health needs of looked after children (including children in foster or kinship care, residential care, or secure units) are a priority area requiring support (NICE, 2021). A great potential for parenting programmes to help address this need whilst also supporting those working as foster carers, who are vulnerable to high levels of stress and burnout (Bridger *et al.*, 2020), has been identified (Kemmis-Riggs and McAloon, 2020; Dorsey *et al*., 2008). Although there are a number of promising programmes which have been adapted or developed specifically for foster carers, to date there is a lack of strong evidence for their effectiveness in the U.K. (NICE, 2021). The RFP is one such intervention which has shown promising results in two feasibility and pilot evaluation studies (Midgley *et al*., 2019, 2021), and is currently being evaluated as part of a large-scale randomised controlled trial (Midgley *et al*., 2021b). For the findings of such trials to be meaningful, it is essential to have a reliable method of assessing whether the intervention has been delivered in a way that can be considered 'on model'. Assessment frameworks need not only to demonstrate adequate psychometric properties, but also be practical to use, especially if they are to be used at a large scale. When fidelity assessments are both reliable and practical to use, they are also important as part of developing a model of training and accreditation for facilitators (Martin *et al*. 2021).

This study suggests that the RFFR, when used as a summary score to assess overall fidelity for the RFP, is both reliable and practical to use. Future studies will need to examine not only levels of fidelity, but also the relationship between programme fidelity and intervention outcomes.

References

Abidin, R. (2012), *Parenting Stress Index: PSI-4; Professional Manual,* PAR.

Bergström, M., Cederblad, M., Håkansson, K., *et al*. (2020), "Interventions in foster family care: a systematic review", *Research on Social Work Practice,* Vol. 30 No.1, pp.3-18. doi:10.1177/1049731519832101

Bond, G. R., Evans, L., Salyers, M. P., Williams, J. and Kim, H. W. (2000), "Measurement of fidelity in psychiatric rehabilitation", *Mental Health Services Research*, Vol. 2 No.2, pp.75–87. https://doi.org/10.1023/A:1010153020697

Borrelli, B. (2011), "The assessment, monitoring, and enhancement of treatment fidelity in public health clinical trials", *Journal of Public Health Dentistry*, Vol. 71, pp.52-S63.

Breitenstein, S. M., Fogg, L., Garvey, C., Hill, C., Resnick, B. and Gross, D. (2010), "Measuring implementation fidelity in a community-based parenting intervention", *Nursing Research*, Vol. 59 No. 3, pp.158–165. https://doi.org/10.1097/NNR.0b013e3181dbb2e2

Bridger, K. M., Binder, J. F. and Kellezi, B. (2020), "Secondary traumatic stress in foster carers: risk factors and implications for intervention", *Journal of Child and Family Studies*, Vol. 29 No. 2, pp.482-492.

Chen, H. and Chen, H. (1990), *Theory-driven evaluations*, Sage, London.

Dorsey, S., Farmer, E. M., Barth, R. P., Greene, K. M., Reid, J. and Landsverk, J. (2008), "Current status and evidence base of training for foster and treatment foster parents", *Children and Youth Services Review*, Vol. 30 No. 12, pp.1403-1416.

Durlak, J. A. and DuPre, E. P. (2008), "Implementation matters: a review of research on the influence of implementation on program outcomes and the factors affecting implementation", *American Journal of Community Psychology*, Vol. 41 No. 3–4, pp.327–350. https://doi.org/10.1007/s10464-008-9165-0

Eames, C., Daley, D., Hutchings, J., Whitaker, C. J., Jones, K., Hughes, J. C. and Bywater, T. (2009), "Treatment fidelity as a predictor of behaviour change in parents attending group-based parent training", *Child: Care, Health and Development*, Vol. 35 No. 5, pp.603–612. https://doi.org/10.1111/J.1365-2214.2009.00975.X

Ellenberg, J. H. (1994), "Selection bias in observational and experimental studies", *Statistics in Medicine,* Vol. 13 No. 5–7, pp.557–567.

Elliott, D. S. and Mihalic, S. (2004), "Issues in disseminating and replicating effective prevention programs", *Prevention Science*, Vol. 5 No. 1, pp.47-53.

Gardner, F. (2000), "Methodological issues in the direct observation of parent–child interaction: do observational findings reflect the natural behavior of participants?", *Clinical Child and Family Psychology Review*, Vol. 3 No. 3, pp.185-198.

Goulet, M., Archambault, I., Janosz, M. and Christenson, S. L. (2018), "Evaluating the implementation of Check & Connect in various school settings: is intervention fidelity necessarily associated with positive outcomes?", *Evaluation and Program Planning,* Vol. 68, pp.34-46. https://doi.org/10.1016/j.evalprogplan.2018.02.004

Hallgren, K. A. (2012), "Computing inter-rater reliability for observational data: an overview and tutorial", *Tutorials in Quantitative Methods for Psychology*, Vol. 8 No.1, p.23.

Hohmann, A. A. and Katherine Shear, M. (2002), "Community-based intervention research: coping with the "noise" of real life in study design", *American Journal of Psychiatry*, Vol. 159 No. 2, pp.201–207. https://doi.org/10.1176/APPI.AJP.159.2.201

Job, A. K., Ehrenberg, D., Hilpert, P., Reindl, V., Lohaus, A., Konrad, K. and Heinrichs, N. (2022), "Taking care Triple P for foster parents with young children in foster care: results of a 1-year randomized trial", *Journal of Interpersonal Violence*, Vol. 37 No. 1-2, pp.322-348.

Kam, C. M., Greenberg, M. T. and Walls, C. T. (2003), "Examining the role of implementation quality in school-based prevention using the PATHS curriculum", *Prevention Science*, Vol. 4 No. 1.

Karterud, S., Pedersen, G., Engen, M., Johansen, M. S., Johansson, P. N., Schlüter, C., Urnes, O., Wilberg, T., & Bateman, A. W. (2013). The MBT Adherence and Competence Scale (MBT-ACS): development, structure and reliability. *Psychotherapy Research*, 23(6), 705–717. https://doi.org/10.1080/10503307.2012.708795

Kazdin, A. E. (1982), Observer effects: Reactivity of direct observation, *New Directions for Methodology of Social & Behavioral Science*.

Kemmis-Riggs, J. and McAloon, J. (2020), "A narrative review of the needs of children in foster and kinship care: informing a research agenda", *Behaviour Change*, pp.1-10.

Kenney, K. (2022). *"MBESS: the MBESS R package". Version 4.9.1*. Available: https://CRAN.R-project.org/package=MBESS

Lindley, L. C. and Slayter, E. M. (2019), "End-of-life trends and patterns among children in the US foster care system: 2005–2015", *Death Studies*, Vol. 43 No. 4, pp.248–259.

Lotty, M., Bantry-White, E. and Dunn-Galvin, A. (2020), "The experiences of foster carers and facilitators of Fostering Connections: The Trauma-informed Foster Care Program: a process study", *Children and Youth Services Review*, Vol. 119, pp. 105516.

Lorencatto, F., West, R., Christopherson, C. and Michie, S. (2013), "Assessing fidelity of delivery of smoking cessation behavioural support in practice", *Implementation Science: IS*, Vol. 8 No.1. https://doi.org/10.1186/1748-5908-8-40

Maaskant, A. M., van Rooij, F. B., Overbeek, G. J., Oort, F. J. and Hermanns, J. M. (2016), "Parent training in foster families with children with behavior problems: follow-up results from a randomized controlled trial", *Children and Youth Services Review*, Vol. 24, pp.84-94.

Martin, M., Steele, B., Lachman, J. M. and 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*, Vol. 24 No. 4, pp.834–853. https://doi.org/10.1007/s10567-021-00350-8

McDaniel, B., Braiden, H. J., Onyekwelu, J., Murphy, M. and Regan, H. (2011), "Investigating the effectiveness of the incredible years basic parenting programme for foster carers in Northern Ireland", *Child Care in Practice*, Vol. 17 No.1, pp. 55-67.

McDonald, R. P. (1999), *Test Theory: A Unified Treatment*, Erlbaum, Mahwah, NJ.

McGraw, K. O . and Wong, S. P. (1996), "Forming inferences about some intraclass correlation coefficients", *Psychological Methods,* Vol. 1, pp. 30–46. doi:10.1037//1082-989X.1.1.30

Midgley, N., Cirasola, A., Austerberry, C., Ranzato, E., West, G., Martin, P., Redfern, S., Cotmore, R. and Park, T. (2019), "Supporting foster carers to meet the needs of looked after children: a feasibility and pilot evaluation of the Reflective Fostering Programme", *Developmental Child Welfare*, Vol. 1 No. 1, pp.41–60. https://doi.org/10.1177/2516103218817550

Midgley, N., Sprecher, E. A., Cirasola, A., Redfern, S., Pursch, B., Smith, C., Douglas, S., and Martin, P. (2021a), "The Reflective Fostering Programme: evaluating the intervention co-delivered by social work professionals and foster carers", *Journal of Children's Services*, Vol. 16 No. 2, pp.159–174. https://doi.org/10.1108/JCS-11-2020-0074

Midgley, N., Irvine, K., Rider, B. *et al*. (2021b), "The Reflective Fostering Programme—improving the wellbeing of children in care through a group intervention for foster carers: a randomised controlled trial", *Trials,* Vol. 22, p. 841. https://doi.org/10.1186/s13063-021-05739-y

Mokkink, L. B., Terwee, C. B., Patrick, D. L., *et al.* (2010a), "The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes", *Journal of Clinical Epidemiology*, Vol. 63 No.7, pp,737-745.

Mokkink, L. B., Terwee, C. B., Patrick, D. L., *et al.* (2010), "The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study", *Quality of Life Research*, Vol. 19, No. 4, pp.539-549.

Moody, G., Coulman, E., Brookes-Howell, L., *et al.* (2020). "A pragmatic randomised controlled trial of the fostering changes programme", *Child Abuse & Neglect*, Vol. 108, pp. 104646.

Moore, G. F., Audrey, S., Barker, M., Bond, L., Bonell, C., Hardeman, W., Moore, L., O'Cathain, A., Tinati, T., Wight, D. and Baird, J. (2015), "Process evaluation of complex interventions: Medical Research Council guidance", *British Medical Journal*, p.350. https://doi.org/10.1136/BMJ.H1258

Mowbray, C. (2003), "Fidelity criteria: development, measurement, and validation," *The American Journal of Evaluation*, Vol. 24 No. 3, pp.315–340. https://doi.org/10.1016/S1098-2140(03)00057-2

Mowbray, C. T., Holter, M. C., Teague, G. B. and Bybee, D. (2003), "Fidelity criteria: development, measurement, and validation", *American Journal of Evaluation,* Vol. 24 No. 3, pp.315–340.

Muris, P., Meesters, C. and van den Berg, F. (2003), "The Strengths and Difficulties Questionnaire (SDQ)", *European Child & Adolescent Psychiatry 2003*, Vol. 12 No. 1, pp.1–8. https://doi.org/10.1007/S00787-003-0298-2

NICE (2010), *Public health guidance: promoting the quality of life of looked-after children and young people*. Available at https://www.scie.org.uk/publications/guides/guide40/files/PH28Guidance.pdf (accessed 12th January 2022).

NICE (2021), *NICE guidance: looked-after children and young people*. Available at: https://www.nice.org.uk/guidance/ng205/chapter/Context (accessed 12th January 2022).

Oakley, A., Strange, V., Bonell, C., Allen, E. and Stephenson, J. (2006), "Process evaluation in randomised controlled trials of complex interventions", *BMJ (Clinical Research Ed.)*, Vol. 332 No. 7538, pp.413–416. https://doi.org/10.1136/BMJ.332.7538.413

Ogrodniczuk, J. S. and Piper, W. E. (1999), "Measuring therapist technique in psychodynamic psychotherapies: development and use of a new scale", *The Journal of Psychotherapy Practice and Research*, Vol. 8 No. 2, p.142.

Paulson, R. I., Post, R. L., Herincks, H. A. and Risser, P. (2002), "Beyond components: using fidelity scales to measure and assure choice in program implementation and quality assurance", *Community Mental Health Journal*, Vol. 38 No.2, pp.119–128. https://doi.org/10.1023/A:1014591020400

Petrowski, N., Cappa, C. and Gross, P. (2017), "Estimating the number of children in formal alternative care: challenges and results", *Child Abuse & Neglect*, Vol.70, pp.388–398. https://doi.org/10.1016/j.chiabu.2016.11.026

R Core Team (2019), *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. URL https://www.R-project.org/

Redfern, S., Wood, S., Lassri, D., Cirasola, A., West, G., Austerberry, C., Luyten, P., Fonagy, P. and Midgley, N. (2018), "The Reflective Fostering Programme: background and development of a new approach", *Adoption & Fostering*, Vol. 42 No. 3, pp.234–248. https://doi.org/10.1177/0308575918790434

Reise, S. P., Ventura, J., Nuechterlein, K. H., and Kim, K. H. (2005), "An illustration of multilevel factor analysis", *Journal of Personality Assessment*, Vol. 84 No. 2, pp. 126–136. https://doi.org/10.1207/s15327752jpa8402_02

Revelle, W. (2021), "*Psych: procedures for psychological, psychometric, and personality research. R package version 2.1.9*", Available: https://CRAN.R-project.org/package=psych .

Shrout, P. E. and Fleiss, J. L. (1979), "Intraclass correlations: uses in assessing rater reliability", *Psychological Bulletin,* Vol. 86, pp.420–428.

Slade, A. (2005), "Parental reflective functioning: an introduction", *Attachment & Human Development*, Vol. 7 No. 3, pp.269-281.

Terwee, C. B., Bot, S. D., de Boer, M. R., *et al.* (2007), "Quality criteria were proposed for measurement properties of health status questionnaires", *Journal of Clinical Epidemiology*, Vol. 60 No. 1, pp.34-42.

Toomey, E., Matthews, J. and Hurley, D. A. (2017), "Using mixed methods to assess fidelity of delivery and its influencing factors in a complex self-management intervention for people with osteoarthritis and low back pain", *BMJ Open*, Vol. 7 No. 8. https://doi.org/10.1136/BMJOPEN-2016-015452

Turner, W. and Macdonald, G. (2011), "Treatment foster care for improving outcomes in children and young people: a systematic review", *Research on Social Work Practice*, Vol. 21 No. 5, pp.501-527.

Vanschoonlandt, F., Vanderfaeillie, J., van Holen, F. and de Maeyer, S. (2012), "Development of an intervention for foster parents of young foster children with externalizing behavior: theoretical basis and program description", *Clinical Child and Family Psychology Review*, Vol. 15 No. 4, pp.330–344. https://doi.org/10.1007/s10567-012-0123-x

de Vet, R., Lako, D. A., Beijersbergen, M. D., van den Dries, L., Conover, S., van Hemert, A. M., Herman, D. B. and Wolf, J. R. (2017), "Critical Time Intervention for People Leaving

Shelters in the Netherlands: Assessing Fidelity and Exploring Facilitators and Barriers", *Administration and policy in mental health*, Vol. 44 No.1, pp.67–80. https://doi.org/10.1007/s10488-015-0699-9

Walton, H., Spector, A., Tombor, I. and Michie, S. (2017), "Measures of fidelity of delivery of, and engagement with, complex, face-to-face health behaviour change interventions: A systematic review of measure quality", *British Journal of Health Psychology*, Vol. 22 No. 4, pp.872–903. https://doi.org/10.1111/BJHP.12260

Weck, F., Bohn, C., Ginzburg, D. M. and Stangier, U. (2011), "Assessment of adherence and competence in cognitive therapy: comparing session segments with entire sessions", *Psychotherapy Research: Journal of the Society for Psychotherapy Research*, Vol. 21 No.6, pp.658–669. https://doi.org/10.1080/10503307.2011.602751

Weck, F., Grikscheit, F., Höfling, V. and Stangier, U. (2014), "Assessing treatment integrity in cognitive-behavioral therapy: Comparing session segments with entire sessions", *Behavior Therapy*, Vol. 45 No. 4, pp.541-552.

Table I: Items of the RFFR.

CONTENT
1. Consistent and explicit focus on the **AIMS/THEMES OF THE SESSION**
2. Use of **MIND CHECKS**
PROCESS AND GROUP DYNAMICS
3. Noticing and naming changes in **EMOTIONAL TEMPERATURE**
4. Noticing and naming **EFFECTIVE MENTALIZING**
5. Noticing, naming and interrupting **INEFFECTIVE OR NON-MENTALIZING**
6. **COOPERATION between co-facilitators**
7. Managing **GROUP DYNAMICS and boundaries**
PROFESSIONAL APP (facilitators take a mentalizing stance)
8. **ATTENTION, CURIOSITY AND INTEREST**
9. **NOT KNOWING STANCE**, avoiding taking an expert position
10. **PROVIDING EMPATHY and validation** of experiences in the group
CARER APP (facilitators promote the use of empathy and perspective-taking)
11. **PROMOTING EMPATHY** towards mental and emotional states of others
12. **PROMOTING PERSPECTIVE-TAKING**
CARER MAP (facilitators promote carer's self-mentalizing)
13. **PROMOTE ATTENTION AND INTEREST IN MENTAL** and emotional states **of SELF AS A CARER** and possible causes and triggers of these
OVERALL SCORE
14. **OVERALL RATING** for the session.

Table II: Descriptive Statistics of RFFR item ratings, by rater and overall

	Overall			Rater A			Rater B			Rater C			Rater D		
Item	Mean	SD	Miss	Mean	SD	Miss	Mean	SD	Miss	Mean	SD	Miss	Mean	SD	Miss
1	3.39	0.97	0	3.50	0.95	0	3.25	0.72	0	3.20	1.06	0	3.60	1.14	0
2	2.79	0.97	27	2.50	1.20	12	2.93	0.80	5	2.67	0.82	5	2.93	1.16	5
3	2.44	0.90	0	2.35	0.99	0	2.70	0.57	0	2.25	0.55	0	2.45	1.28	0
4	2.61	1.09	1	2.63	0.90	1	2.85	1.09	0	2.45	1.15	0	2.50	1.24	0
5	2.00	1.05	1	2.16	0.83	1	1.85	1.04	0	1.80	1.06	0	2.20	1.24	0
6	2.73	0.80	0	3.35	0.59	0	2.45	0.69	0	2.55	0.76	0	2.55	0.83	0
7	3.09	0.98	0	3.25	0.85	0	3.00	0.73	0	3.00	0.97	0	3.10	1.33	0
8	3.10	1.03	0	3.35	0.81	0	3.05	1.00	0	3.15	0.93	0	2.85	1.31	0
9	2.61	0.91	0	3.10	1.02	0	2.60	0.68	0	2.35	0.81	0	2.40	0.94	0
10	3.23	0.93	0	3.55	0.60	0	3.30	0.80	0	3.15	1.09	0	2.90	1.07	0
11	2.56	0.94	1	2.42	0.69	1	2.75	0.79	0	2.40	1.19	0	2.65	1.04	0
12	2.62	0.95	0	2.40	0.75	0	2.80	0.89	0	3.00	1.08	0	2.30	0.92	0
13	2.25	1.03	1	2.68	0.89	1	2.30	0.98	0	1.90	1.17	0	2.15	0.99	0
14	2.73	0.89	1	3.05	0.83	0	2.70	0.80	0	2.75	0.79	0	2.42	1.07	1
Summary Score	38.15	9.36	-	40.62	7.61	-	38.47	8.60	-	36.54	8.75	-	36.95	12.06	-

Notes: Miss: Number of missing ratings. Where missing values occurred in a clip set, the summary score was prorated using the average of the available ratings to replace missing ratings.

Table III: RFFR items: pooled within-rater Pearson correlations

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	-													
2	0.48	-												
3	0.25	0.45	-											
4	0.41	0.59	0.52	-										
5	0.46	0.50	0.45	0.64	-									
6	0.27	0.37	0.42	0.48	0.42	-								
7	0.60	0.46	0.33	0.48	0.45	0.39	-							
8	0.48	0.67	0.50	0.68	0.64	0.48	0.61	-						
9	0.12	0.42	0.24	0.26	0.15	0.20	0.25	0.54	-					
10	0.05	0.43	0.32	0.32	0.22	0.40	0.28	0.59	0.61	-				
11	0.32	0.31	0.40	0.38	0.38	0.38	0.40	0.62	0.48	0.46	-			
12	0.30	0.48	0.39	0.45	0.39	0.29	0.48	0.71	0.59	0.60	0.74	-		
13	0.45	0.51	0.43	0.52	0.46	0.40	0.42	0.63	0.35	0.30	0.55	0.53	-	
14	0.55	0.72	0.61	0.74	0.64	0.54	0.65	0.86	0.47	0.51	0.65	0.69	0.73	-

Note: Mean replacement was used to impute missing ratings. 50 correlations are based on complete ratings, 28 use a single imputed rating, and 2 use two imputed ratings.

Table IV. Loadings from a multilevel factor analysis of RFFR ratings

		One factor	Two factors	
			Factor 1	Factor 2
Item	1	**0.551**	**0.731**	–0.203
	2	**0.733**	**0.654**	0.134
	3	**0.607**	**0.566**	0.077
	4	**0.748**	**0.820**	–0.054
	5	**0.669**	**0.804**	–0.135
	6	**0.552**	**0.519**	0.066
	7	**0.660**	**0.645**	0.052
	8	**0.905**	**0.636**	**0.390**
	9	**0.515**	–0.088	**0.804**
	10	**0.555**	–0.015	**0.763**
	11	**0.676**	0.254	**0.578**
	12	**0.732**	0.207	**0.720**
	13	**0.729**	**0.621**	0.169
	14	**0.965**	**0.808**	0.242
Proportion of variance explained	Per factor	0.486	0.346	0.172
	Total	0.486	0.518	
Between-factor correlation		-	0.557	

*Note: N = 80. Factor analysis was conducted on the pooled within-rater covariance matrix. Loadings with absolute values > 0.3 are highlighted **bold**. Mean replacement was used to impute missing values.*

Table V: Estimated intraclass correlations (ICC(3,1)) for RFFR items and RFFR summary score

Item	ICC	(95 % CI)	Percentage agreement
1	0.50	(0.27, 0.72)	42.5%
2	0.43	(0.21, 0.67)	43.5%
3	0.32	(0.10, 0.58)	43.3%
4	0.64	(0.44, 0.81)	41.9%
5	0.49	(0.26, 0.71)	39.3%
6	0.49	(0.27, 0.72)	47.5%
7	0.28	(0.07, 0.55)	41.7%
8	0.61	(0.40, 0.80)	43.3%
9	0.50	(0.28, 0.72)	37.5%
10	0.51	(0.28, 0.73)	40.8%
11	0.34	(0.12, 0.60)	33.3%
12	0.59	(0.37, 0.78)	38.3%
13	0.53	(0.30, 0.74)	35.9%
14	0.71	(0.53, 0.86)	57.3%
Summary Score	0.74	(0.57, 0.87)	--

Note: N = 80 per item, using 4 raters on 20 clip sets each. The summary score was prorated where item ratings were missing. Overall percentage agreement across all 14 items: 41.8 %.

Table VI: RFFR Summary Scores for 20 group session videos, by rater and overall

Clip Set Number	Rater A	B	C	D	Mean
10	32.3	25.8	22.6	23.7	26.1
20	31.2	28.0	30.0	20.0	27.3
11	33.4	28.0	27.0	27.0	28.8
2	33.0	34.0	31.0	20.0	29.5
5	37.3	32.0	33.0	23.0	31.3
8	34.0	34.0	28.0	32.0	32.0
15	36.6	34.5	29.1	28.0	32.0
17	38.8	31.2	28.0	34.5	33.1
13	45.0	30.0	30.0	31.2	34.1
18	24.8	38.0	36.0	43.0	35.4
19	41.0	33.0	36.0	33.0	35.8
4	46.3	41.0	38.0	32.0	39.3
1	39.8	43.0	41.0	48.0	43.0
3	51.0	42.0	39.0	44.0	44.0
12	50.2	48.0	44.0	35.0	44.3
14	46.3	44.2	38.8	50.6	45.0
7	48.5	50.0	46.0	48.0	48.1
9	45.2	48.0	53.0	56.0	50.6
6	47.0	52.0	53.0	54.0	51.5
16	50.6	52.8	47.4	56.0	51.7

Note: Clip sets are ordered by average rating in ascending order.

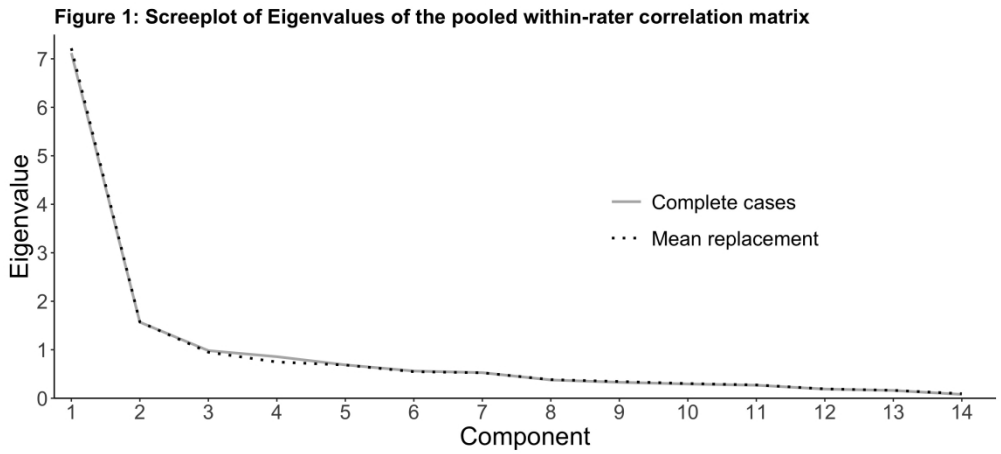

Figure I: Screeplot (PCA on pooled within-rater correlations)

1164x529mm (72 x 72 DPI)