Scandinavian Journal of Psychology, 2014, 55, 65–71 DOI: 10.1111/sjop.12086

Personality and Social Psychology

Observations of families in structured interactions: Parenting therapists provide reliable ratings of mothers' parenting

BENT STORÅ,[1] KNUT A. HAGTVET[2] and SONJA HEYERDAHL[3]

[1]*Department of Child and Adolescent Mental Health, Sorlandet Hospital and Centre for Child and Adolescent Mental Health, Eastern and Southern Norway, Norway*
[2]*Department of Psychology, University of Oslo, Oslo, Norway*
[3]*Centre for Child and Adolescent Mental Health, Eastern and Southern Norway (RBUP Oslo), Oslo, Norway*

Storå, B., Hagtvet, K. A. & Heyerdahl, S. (2014). Observations of families in structured interactions: Parenting therapists provide reliable ratings of mothers' parenting. *Scandinavian Journal of Psychology* 55, 65–71.

The reliability of observations of parenting by parenting therapists was assessed. An important predictor of externalizing behavior in children is quality of parenting. Data were videotapes of structured interactions in families with a child age 8–12 years referred to the evidence based Parent Management Training Oregon (PMTO) treatment program for child behavior problems. The therapists had clinical PMTO training but no training in systematic observation. PMTO observational coders with specific coder training were included as a reference for the therapists. Five therapists and two coders observed videotapes of 10 families and performed global evaluations of mothers' parenting skills. They used the coder's impression measure used in PMTO research. Scores were analyzed in a generalizability theory framework for the two groups of observers separately. Both observer types reliably rank-ordered the mothers and assessed the level of parenting skills. PMTO therapists without coder training provided reliable ratings of parenting constructs relevant to the clinical PMTO program in a manner comparable to that of the trained reference coders.

Key words: Global observation, therapist observers, family interactions, generalizability theory.

Bent Storå, Department of Child and Adolescent Mental Health, Sorlandet Hospital and Centre for Child and Adolescent Mental Health, Eastern and Southern Norway, Service Box 416, N-4604 Kristiansand, Norway. Tel: +4791735711; fax: +4738076257; e-mail: bent.stora@sshf.no

INTRODUCTION

Observations of families have the advantage over self reports and interviews that the information is not filtered through the perceptions of the involved parents (Couteur & Gardner, 2008). But observer bias can be a problem that can influence rater reliability since observations are filtered through the observer (Chafouleas, Christ, Riley-Tilman, Briesch & Chanese, 2007). The therapist is the observer in clinical practice, whereas in research projects observation is often performed by specially trained coders. Therapists are natural participants in therapeutic situations and possess experiences that are relevant to clinical research. There seems to be a paucity of knowledge concerning therapists in the role of observers of parent-child interactions. The present study assessed the reliability of global observations of parenting by Parent Management Therapy Oregon (PMTO; Forgatch & DeGarmo, 2002; Forgatch & Patterson, 2010; Ogden & Hagen, 2008). Families were observed in the setting of a parent-child interaction, which was structured to elicit behaviors that represent clinically salient domains.

PMTO is a well-established evidence based treatment program for children with disruptive behavior (Eyberg, Nelson and Boggs, 2008). Caregivers' parenting skills have been identified as an important mediator of child behavior problems along with deviant peer relations (Eddy & Chamberlain, 2000). There are five principal parenting constructs that are included in the PMTO program: Effective *discipline* and limit setting discourage deviant behavior through the appropriate and contingent use of mild sanctions. Such parenting provides the child with clear boundaries for acceptable behavior. *Skill encouragement* promotes competence through positive contingencies. *Positive involvement* reflects how parents demonstrate interest in, attention to and care for their child. *Problem-solving* helps family members negotiate disagreements, establish house rules, and specify consequences for following or violating rules. *Monitoring* prevents youngsters from involvement in risky activities and monitoring reflects parental tracking of the child's whereabouts. (Forgatch & DeGarmo, 2002; Ogden & Hagen, 2008). A consistent finding in PMTO research is that effective *discipline* is a mediator for follow-up improvements in several child behavior problem domains (Hagen, Ogden & Bjørnebekk, 2011; Ogden & Hagen, 2008; Patterson & Forgatch, 1995). Observations of family interaction have been essential in the development and evaluation of the treatment program. The participating families are videotaped in structured interactions before therapy, at treatment termination and at follow up in PMTO research. In structured observation the family is observed in contexts that involve the manipulation of environmental conditions to sample target behaviors (Mori & Armendariz, 2001; Roberts & Hope, 2001). The situation is structured in different ways, depending on what types of behavior the researcher wants to elicit. Scenarios targeting child externalizing problems typically focus on eliciting cooperative behavior or family conflicts (Mori & Armendariz, 2001; Roberts & Hope, 2001). The assumption that the behavior of families interacting in a laboratory setting is analogous to family interactions as they occur in everyday life has been questioned due to the artificiality of such observational contexts. Structuring of the observations is however merited for research on observable behavior in well controlled studies such as the PMTO program. (Gardner, 2000).

Coding of interactions is a procedure that can be performed at various levels of analysis (Lindahl, 2001). In detailed (microanalytic) observational methods second-to-second interactions of small units of behavior are coded and in global methods large coding units that require the coders to apply global judgment are coded. Microanalytic methods may be more robust against observer biases and more sensitive to change in clinical intervention studies relative to global ratings (Snyder, Reid, Stoolmiller, Howe, Brown & Dagne, 2006). However; Patterson and Forgatch (1995) reported that global observations of parenting predicted long-term adjustment better than change in child adjustment from baseline to post treatment. Global observations are indirect in the sense that the observers are required to make inferences regarding the fit between observed behaviors and latent constructs using items that are indicators of the construct. For certain domains untrained observers might possess a form of "intuitive knowledge" (Waldinger, Schulz, Hauser, Allen & Crowell, 2004). Shared personal and cultural experiences facilitated the ratings of parenting skills by untrained coders and resulted in high concordance with expert ratings, whereas the lack of relevant experience resulted in moderate concordance with expert ratings on maternal sensitivity (Baker, Messinger, Ekas, Lindahl & Brewster, 2010). Thus, global observation might be especially amenable to minimally trained coders (Lorber, 2006) and relevant experience might constitute a form of observer training (Chafouleas *et al*., 2007).

Generalizability (G) theory (Brennan 2001; Cronbach, Gleser, Nanda, Rajaratnam, 1972) is a flexible system that allows the assessment of the degree to which a given set of observations generalizes to a more extensive set of observations. The G-theory framework gives the opportunity to disentangle several variance components within a measurement design in one analysis in contrast to classical test theory. From the relative size of these components inferences can be made about the generalizability of scores and the dependability of the different facets of observation. (Brennan, 2001). G-theory is uniquely suited to reliability studies of observational data for behavior that is prompted by a predetermined set of environmental stimuli, particularly in the assessment of observer influence on the results of the observation (Hintze & Matthews, 2004). The measurement design in the present study consists of more than one facet justifying the application of G-theory as a framework for estimating reliability (Brennan, 2001). Two types of studies are assumed in G-theory: In G-studies, variance due to raters, items and other facets of observation is partitioned into variance components that, in turn, are used in D-studies to assess generalizability coefficients that are tailored to the sources of error in the measurement design. In D-studies, changes in the generalizability coefficients may be assessed if the researcher decides to change the number of conditions within facets of observations, such as the number of raters and the number of items.

Parenting therapists and trained coders were both used as observers of families engaged in predetermined tasks in the present study. Our expectation was that observation of parenting at such a subjective and global level would be intuitive to PMTO therapists because of their clinical experiences, and as a result, they would be able to generate inferences regarding parenting practices in a reliable manner.

Aims

The first aim was to assess the reliability for PMTO-therapists' ratings of each of the five parenting practices: *discipline*, *positive involvement*, *problem-solving*, *skill encouragement*, and *monitoring*. More specifically the aim was to assess the rank-ordering and level of parenting skills based on G-theory. The results for the PMTO therapists are compared to those for trained coders. The second aim was to estimate the number of PMTO-therapist raters and trained coders needed to obtain reliable scores for each parenting practice.

METHOD

Participants

The 10 participating families were selected from a pool of 112 families that participated in a randomized control trial of PMTO in Norway (Ogden & Hagen, 2008). The families participated in structured family interaction tasks as part of the assessment procedure. A team of trained coders rated videotapes from the tasks. From the videotapes, 20% were randomly selected for observation by two raters for reliability assessment. The 10 families that were included in the present study were arbitrarily chosen from the 20% of the families that were rated by the same two coders. The data were collected in 2005 from families living in Norway and who were referred for child conduct problems. These 10 families with children above the age of 8 to 12 years were also observed by five PMTO therapists. The child was a boy in five families and a girl in the other five. The father was present in half and the mother was present in all of the observations.

PMTO therapists. Five PMTO therapists were observers; three females and two men, aged 38 to 54. Due to their PMTO license the therapists are considered to be representative of other licensed PMTO therapists. Along with the videotapes they received the Coder's Impression (CI) measure (Forgatch, Knutson & Mayne, 1992), scoring sheets, and a written instruction comparable to the instructions to the trained coders. None of the PMTO therapists had knowledge of the families. To achieve the PMTO license the candidates participated in group supervision for about 200 hours and they were required to complete at least five family treatments and demonstrate an acceptable level of fidelity to the PMTO program manual (Ogden, Forgatch, Askeland, Patterson & Bullock, 2005).

Coders. The two coders were both female, aged 22 and 25 and were part of a team of coders that were trained to reliability in the PMTO studies in Norway (Ogden & Hagen, 2008). The coders observed the family interactions using two different observational formats: the macroanalytic CI measure (Forgatch *et al*., 1992) which is utilized in the present study and the microanalytic Family and Peer Process Code (FPPC) developed by Stubbs, Crosby, Forgatch & Capaldi (1998). The coders completed the CI measure after finishing the microanalytic coding. The training for the macrosocial CI measure consisted of the observation team watching videotapes together of sample films of families in structured interactions. They were considered reliable when they had 80% agreement with an expert coder. Reliability was reached in 7–8 hours of watching sample videotapes. We have reported acceptable generalizability for the global observations of the coders in the coding team in another publication (Storå, Hagtvet & Heyerdahl, 2012) which is an indication that the coder training was sufficient.

Procedures

The structured interaction tasks took place in a laboratory or clinic and lasted for 30 minutes. All of the families engaged in a set of tasks as directed by the test administrator. In the first task the family was instructed to spend five minutes planning something nice to do together

during the next week. The second and third tasks were 10-minute problem-solving tasks. The parents chose the issue in the second task and the child chose the issue in the third task. They were instructed to choose issues that often create conflicts in families from the Issues Checklist; for example, chores, bed time, TV and computer time (Prinz, Foster, Kent & O'Leary, 1979). In the fourth task, the family was instructed to discuss the quality of their interaction for five minutes. The observers stopped the film and rated the CI items after each of the four structured interaction tasks and after observing all tasks. The PMTO study was approved by the Regional Ethical Committee for Medical Research Ethics, Southern Norway, and the Norwegian Data Inspectorate.

Measures

Global observations using the CI observational measure (Forgatch *et al*., 1992) have been used in PMTO research since the 1980s to allow coders trained in microanalytic interaction coding to rate global impressions of family interactions (Forgatch & DeGarmo, 2002; Ogden & Hagen, 2008; Patterson 1982). The CI measure consists of items representing indicators of the five principal parenting constructs of the PMTO program. Furthermore, the measure has been found to be sensitive to change and to have good convergent and predictive validity (DeGarmo, Patterson & Forgatch, 2004). The CI measure consists of well-defined Likert scale items describing the quality, content and characteristics of the interaction, with an emphasis on parenting practices (Bullard, Wachlarowicz, DeLeeuw *et al*., 2010). The CI measure was translated for the PMTO randomized controlled effectiveness study in Norway (Ogden & Hagen, 2008) by bilingual members of the research group, who were supervised by reference persons that were familiar with the instruments from previous research. An 88-item version with items pertaining to the mothers was used in the present study.

Examples of items: *Discipline* (13 items): "Discipline style is overly strict;" *Skill encouragement* (4 items): "Skillfully prompted the youngster during the task as necessary;" *Problem-solving* (31items): "Showed willingness to discuss ideas suggested by others;" *Positive involvement* (32 items): "The quality of the relationship between the mother and child was excellent;" and *Monitoring* (11 items): "The mother gathered information from the youngster about activities/friends in an appropriate manner" (e.g., direct, straightforward, interested, pleasant).

Data analysis procedures

Missing data. Missing occurred when the coder had not rated an item. Three items were excluded from the analyses because of an unacceptably high level of missing data on the *monitoring* scale, resulting in an eight item scale. There was no missing data for *discipline*, and missing data on one item for the other parenting practices. For each family missing values were replaced by the mean value taken across the items with non-missing values for the actual scale.

Generalizability theory applied to the present observational measurement design

Two facets of observations were applied in the measurement design: raters (r) and items (i). Variation due to mothers (m) is used to define the object of measurement.

The present application of G-theory rests on the assumption that any observed behavioral indicator belongs to a *representative* sample of behavioral indicators from a hypothetical universe of all similar behavioral indicators. Rater is treated as so-called random facet: The present sample of PMTO raters is assumed to be exchangeable with other samples from the same universe of trained PMTO raters. Items (i) are treated as a random facet when the items are assumed to represent other similar items. Items (i) are also treated as a fixed facet of observation; in this case generalizations are not made beyond the items in the CI measure. This second so called D-study provides a way to define an inter-rater reliability coefficient when generalizing over raters only (Brennan, 2001). The present sample of mothers is assumed to constitute a sample of a population of mothers of children referred for behavior problems.

Father was present in half of the observed families and a facet of observation designating whether the father was present or absent should therefore be included. First, G-theory analyses were conducted with a father facet included. No variance could be attributed to the father being present or absent for any of the five parenting practices for either rater type and we decided to reduce the complexity of the model by excluding the father facet.

A mother by rater by item (mri) design is required when the analyses are run for each rater type separately, that is, for items (i) and raters (r), with mothers (m) as the object of measurement. Based on this mri measurement design, seven sources of variation can be identified. For an illustration see the Venn diagram (Figure 1) at www.sshf.no/stora. The variance component of mother ($\sigma^2_{\ m}$) represents the extent to which mothers differ from each other. The rater (σ^2_r) component represents inconsistencies in the mean values of the raters across the other facets, that is, in how strict or lenient their ratings are, whereas the item (σ^2_i) component indicates inconsistencies in mean values across items.

The mother by rater (σ^2_{mr}) and mother by items (σ^2_{mi}) components represent the extent to which the rank order of mothers differs across raters and items, respectively. The rater by items (σ^2_{ri}) component represents inconsistencies of rater mean values across items. Because mothers are crossed with items which in turn are crossed with raters the rank order of mothers may vary across combinations of raters and items. This last interaction variance component encompasses unmeasured facets and/or random events (e) and is designated as ($\sigma^2_{mri(e)}$). The relative size of each variance component indicates how much of the total variance can be attributed to each specific source of variation.

The generalizability coefficient ($E\rho^2$) focuses on the *relative standing* of the mothers and the dependability coefficient or the absolute generalizability coefficient referred to as Φ, focuses on the *absolute* scores for the mothers. The generalizability coefficient is relevant when assessing consistency in the relative ranking of mothers across one or more facets, whereas the dependability coefficient is relevant when assessing the consistency in absolute levels of performance across one or more facets independent of the scores of other persons. The absolute coefficient is equal to or less than the relative G-coefficient as more variance terms are included in the error variation. (Brennan, 2001; Hintze & Matthews, 2004; Shavelson & Webb, 1991). The rater (σ^2_r) component, the item (σ^2_i) component and the rater by item (σ^2_{ri}) component affect only the dependability coefficient, not the generalizability coefficient in the present design.

Generalizability and dependability coefficients were first estimated for the complete multi-facet design generalizing across both raters and items. Secondly, redefined D-study estimation formulae assuming raters and items as random and fixed facets, respectively, were applied. See Appendix B and C at www.sshf.no/stora for relevant estimation formulae. This second D-study provides a way to define an inter-rater reliability coefficient when generalizing over raters only, and also to estimate reliability for different numbers of raters. Inter-rater reliability is of particular interest for determining the number of raters that would be needed to obtain a certain level of generalizability and dependability when assessing parental practices.

RESULTS

Table 1 provides descriptive statistics for the five parenting practice subscales. A high score indicates better parenting practices. The coders rated the mothers' parenting significantly higher than the therapists on *discipline* (t (9) = –3.55, $p < 0.01$), and *monitoring* (t (9) = –3.65, $p < 0.05$). The therapists rated the mother significantly higher than the coders on *problem solving* (t (9) = 4.63, $p < 0.01$).

The estimated G-study variance components for the five parenting practices are reported in Table 2. These are also

Table 1. *Summary statistics of subscales of parenting practices as rated by licensed PMTO therapists and coders*

	Discipline 13 items, 1–5 scale Mean (Sd)	Skill Encouragement 4 items, 1–7 scale Mean (Sd)	Positive Involvement 32 items, 1–7 scale Mean (Sd)	Problem solving 31 items, 1–7 scale Mean (Sd)	Monitoring 8 items, 1–7 scale Mean (Sd)
5 licensed PMTO therapists					
Therapist 1	3.48 (0.56)	4.34 (1.34)	4.69 (1.12)	3.48 (1.57)	5.19 (1.00)
Therapist 2	3.28 (0.60)	5.15 (0.94)	4.54 (1.12)	4.30 (1.17)	4.44 (1.28)
Therapist 3	3.19 (0.68)	4.56 (1.49)	4.57 (1.42)	3.49 (1.10)	4.57 (1.24)
Therapist 4	3.16 (0.49)	4.84 (1.00)	4.64 (1.04)	4.05 (1.24)	4.77 (0.80)
Therapist 5	3.07 (0.59)	4.83 (1.00)	4.57 (0.97)	4.01 (1.18)	4.67 (0.70)
Mean	3.23 (0.58)	4.74 (1.15)	4.60 (1.13)	3.86 (1.25)	4.73 (1.00)
Coders					
Coder 1	3.83 (0.53)	4.45 (0.71)	4.56 (1.06)	3.48 (1.08)	5.20 (0.86)
Coder 2	3.82 (0.85)	4.55 (1.62)	4.55 (1.20)	3.54 (1.24)	5.15 (0.45)
Mean	3.83 (0.69)	4.50 (1.17)	4.56 (1.13)	3.51 (1.16)	5.18 (0.66)

Note: N = 10 families. Higher scores indicate better parenting practices.

presented as percentages of the total variance for each subscale to illustrate the relative importance of each component. The variance components of scores from the PMTO therapists and the coders follows a comparable pattern for the parenting practice *discipline*, except that there is substantially more variance caused by items for the coders. There was a comparable variance component structure for the other four parenting scales for both PMTO therapists and coders.

The generalizability coefficients were estimated based on the G-study variance components in Table 2. The estimated relative generalizability coefficients ($E\rho^2$) for the PMTO therapists ranged from 0.87 to 0.93. This finding suggests that the universe score variance explained 87% to 93% of the expected observed variance and is an indication of acceptable generalizability coefficients for the ratings of mothers' parenting. The corresponding estimates of the absolute generalizability coefficients (dependability coefficients; (Φ) varied from 0.80 to 0.93. For the coders, the estimated generalizability coefficients ($E\rho^2$) ranged from 0.58 to 0.83 and the absolute coefficients from (Φ) 0.47 to 0.81. The main reason for higher generalizability coefficients in the therapists group is that there are more raters in that group.

The generalizability coefficients in Table 2 were obtained with the assumption that generalizations are made across both raters and items. Of primary importance in the present study is a focus on the D-study reliability estimation of the ratings of the observers when assuming fixed sets of items, that is, when not generalizing to other sets of items representing the respective parenting constructs. In a series of D-studies, coefficients were estimated for different numbers of raters (Table 3). Coefficients for the two types of raters can be compared when considering the same number of raters, namely, two raters of each type. The relative coefficients for the subscales *discipline*, *skill encouragements* and *monitoring* were somewhat lower for the PMTO therapists than for coders, while the coefficients obtained for *positive involvement* and *problem solving* were considered equal.

The absolute **(Φ)** generalizability coefficient attracts some interest as an inter-rater reliability coefficient. While $E\rho^2$ is reflecting inconsistency with respect to the rank order of mothers' parenting ability Φ is in addition affected by inconsistency due to stringency of raters. When comparing the absolute coefficient for the two types of raters, respectively, the coefficients were considered equal for *positive involvement* and *problem solving*, while the PMTO raters obtained lower Φ for *discipline*, *skill encouragement* and *monitoring* (Table 3).

There is little difference between the relative ($E\rho^2$) and absolute (Φ) generalizability coefficients for either rater type with exception for therapist-rated *discipline* (Table 3). This indicates that there is generally weak contributions of both item and rater variance (Table 2) that impact the absolute error variance of the absolute generalizability coefficient.

Comparing the results in Tables 2 and 3 for the same number of PMTO raters (n'_r = 5), both types of coefficient did not decrease much from the fixed facet condition in Table 3 to the random condition in Table 2: when generalizing beyond the items in each subscale to similar items generalizability remained acceptable for the PMTO therapists. The same trend was not observed for the coders who obtained consistently lower estimates in the random model (Table 2) except for *problem solving*, where the estimates remained the same.

DISCUSSION

The present study assessed the reliability of PMTO therapists and coders that observed videotapes of families in structured interaction utilizing a global observation format that was conceptually relevant to PMTO therapy. The main finding was that PMTO therapists with clinical training provided reliable ratings of family interactions across all of the five principal PMTO constructs. The present sample of PMTO therapists is probably representative for PMTO therapists, as they were certified PMTO therapists. The reliability of the PMTO therapists was at about the same level and with the same pattern of variance components as the coders both in the relative ($E\rho^2$) rank-ordering of the mothers and in their judgments about the mothers' absolute (Φ) level of parenting competency. We are not aware of other studies that utilize parenting therapists as observers.

There are no universally accepted levels of reliability applied to all measurement contexts. There is, however, a convention that a test used for individual purposes should be more reliable than one used for groups or correlation purposes. (McDonald, 1999). A less rigorous level of precision will be needed if the

Table 2. *Estimated G-study variance components, relative generalizability coefficients $E\rho^2$, and dependability coefficients Φ for the random model – mri-design*[a]

	Discipline 13 items		Skill Encouragement 4 items		Positive Involvement 32 items		Problem solving 31 items		Monitoring 8 items	
PMTO Therapists	df	σ^2_α (%)	df	σ^2_α (%)	df	σ^2_α (%)	df	σ^2_α (%)	df	σ^2_α (%)
m	9	0.23 (21.1)	9	0.84 (38.7)	9	1.10 (42.2)	9	1.18 (34.5)	9	0.69 (30.1)
r	4	0.08 (7.3)	4	0.01 (0.5)	4	0	4	0.07 (2.1)	4	0.04 (1.8)
i	12	0.05 (4.6)	3	0	31	0.03 (1.2)	30	0.41 (12.0)	7	0.04 (1.8)
mr	36	0.08 (7.3)	36	0.33 (15.2)	36	0.34 (13.0)	36	0.42 (12.3)	36	0.21 (9.2)
mi	108	0.17 (15.6)	27	0.01 (0.5)	279	0.21 (8.1)	270	0.3 (9.7)	63	0.34 (14.9)
ri	48	0.09 (8.3)	12	0.14 (6.5)	124	0.02 (0.8)	120	0.05 (1.5)	28	0.09 (3.9)
mri(e)	432	0.39 (35.8)	108	0.84 (38.7)	1116	0.91 (34.9)	1,080	0.96 (28.1)	252	0.88 (38.4)
Total Variance		1.09		2.17		2.61		3.42		2.29
$E\rho^2$		0.87		0.88		0.93		0.92		0.87
Φ		0.80		0.88		0.93		0.90		0.85
Coders										
m	9	0.32 (19.8)	9	0.66 (20.1)	9	1.01 (43.4)	9	0.96 (29.2)	9	0.33 (24.1)
r	1	0	1	0	1	0	1	0	1	0
i	12	0.34 (21.0)	3	1.01 (32.0)	31	0.07 (3.0)	30	0.66 (20.1)	7	0.05 (3.7)
mr	9	0.11 (6.8)	9	0.71 (22.5)	9	0.24 (10.3)	9	0.36 (10.9)	9	0.04 (2.9)
mi	108	0.25 (15.4)	27	0.24 (7.6)	279	0.18 (7.7)	270	0.25 (7.6)	63	0.35 (25.6)
ri	12	0.01 (0.6)	3	0	31	0.13 (5.6)	30	0.08 (2.4)	7	0.10 (7.3)
mri(e)	108	0.59 (36.4)	27	0.54 (17.1)	279	0.70 (30.0)	270	0.98 (29.8)	63	0.50 (36.5)
Total Variance		1.62		3.16		2.33		3.29		1.37
$E\rho^2$		0.77		0.58		0.80		0.83		0.78
Φ		0.72		0.47		0.78		0.81		0.76

Notes: m = mothers, r = raters, i= items, e = unmeasured facets that affect the measurement and/or random events. Estimations are based on five and two raters for the PMTO therapists and coders, respectively.[a]Generalizing over raters and items.

Table 3. *Inter-rater reliability coefficients for the mixed model – mri-design[a]. Coefficients are estimated for different numbers of raters*

PMTO Therapists	Discipline 13 items		Skill Encouragement 4 items		Positive Involvement 32 items		Problem solving 31 items		Monitoring 8 items	
	$E\rho^2$	Φ	$E\rho^2$	Φ	$E\rho^2$	Φ	$E\rho^2$	Φ	$E\rho^2$	Φ
$n_r = 5$	0.92	0.86	0.90	0.89	0.93	0.93	0.93	0.91	0.92	0.91
$n_r = 4$	0.90	0.83	0.87	0.86	0.92	0.92	0.91	0.89	0.90	0.89
$n_r = 3$	0.87	0.79	0.84	0.82	0.90	0.90	0.89	0.87	0.87	0.85
$n_r = 2$	0.80	0.70	0.77	0.76	0.86	0.86	0.84	0.82	0.82	0.79
$n_r = 1$	0.69	0.55	0.63	0.61	0.75	0.75	0.76	0.70	0.68	0.66
Coders										
$n_r = 2$	0.86	0.86	0.85	0.85	0.85	0.85	0.83	0.83	0.88	0.87
$n_r = 1$	0.76	0.76	0.74	0.74	0.73	0.73	0.71	0.71	0.79	0.74

Notes: Relative generalizability coefficients Eρ2 and dependability coefficients Φ for different number of raters (n_r).[a]Generalizing over raters as the random facet only, item treated as a fixed facet.

observation is to be used to decide whether a mother has achieved adequate skills on one parenting practice to move on with the program than if the observation is intended for judicial purposes. We propose that the results in Table 3 can be used as an indication of the number of observers needed to achieve a relevant threshold of rater reliability.

Two universe specifications were utilized in the present study. In the first, generalizations were made over both raters and items (Table 2). The second universe formulation was narrower, and generalizations were made over raters only to estimate a type of inter-rater reliability (Table 3). The reliability did not decrease very much when the item facet was treated as a random facet of observation (Table 2) compared with the analyses with items as a fixed facet (Table 3). Generalizability was estimated to be acceptable when items are also considered to be random which adds to the generalizability of the present findings.

We suggest that the main reason for the positive results is that the clinical PMTO training constitutes training in the observation of parenting skills and we do not expect that these positive results can be generalized to clinicians other than PMTO therapists.

Limitations

A concern in the present study is related to the small sample size of mothers, and of therapist and coder observers. The two coders in the present study were part of a team of coders. There are only minor differences between the G-study variance component composition and generalizability coefficients for the observations by the two coders in the present study and results for the coding team that we studied previously (Storå, Hagtvet & Heyerdahl, 2012). We see this as an indication that the results for the coders are representative of the coding team. It may be noted that the applied ANOVA estimation method makes no distributional-form assumptions and G-theory is a practical approach no matter how large the data set may be (Brennan, 2001). Nevertheless, the estimated variance components would have been more stable with increasing person sample size. Therefore larger person samples are solicited in future replications of our study.

It may be noted that influential error represented by inconsistency in mothers' rank order across items, occurred in some of the subscales for both types of rater (Table 2). This may reflect heterogeneity in item content of some of the subscales. This observation may be an issue for future research.

CONCLUSIONS

PMTO therapists provided reliable ratings of parenting based on observations of structured interaction tasks. The global items that were applied are conceptually relevant to PMTO therapy, and PMTO therapists have relevant experiences and knowledge with parenting abilities from their clinical training.

REFERENCES

Baker, J. K., Messinger, D. S., Ekas, N. V., Lindahl, K. M. & Brewster, R. (2010). Nonexpert ratings of family and parent-child interaction. *Journal of Family Psychology*, *24*, 775–778.

Brennan, R. L. (2001). *Generalizability theory: Statistics for social science and public policy*. New York: Springer.

Bullard, L., Wachlarowicz, M., DeLeeuw, J., Snyder, J., Low, S., Forgatch, M. & DeGarmo, D. (2010). Effects of the Oregon Model of Parent Management Training (PMTO) on marital adjustment in new stepfamilies: A randomized trial. *Journal of Family Psychology*, *24*, 485–496.

Chafouleas, S. M., Christ, T. J., Riley-Tillman, T. C., Briesch, A. M. & Chanese, J. A. M. (2007). Generalizability and dependability of direct behavior ratings to assess social behavior of preschoolers. *School Psychology Review*, *36*, 63–79.

Couteur, A. L. & Gardner, F. (2009). Use of structured interviews and observational methods in clinical settings. In M. Rutter, D. V. M. Bishop, D. S. Pine, S. Scott, J. Stevenson, E. Taylor & A. Thapar (Eds.), *Rutter's child and adolescent psychiatry*, (5th edn). Oxford: Blackwell Publishing.

Cronbach, L. J., Gleser, G C., Nanda, H. & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles*. New York: John Wiley.

DeGarmo, D. S., Patterson, G. R & Forgatch, M. S. (2004). How do outcomes in a specified parent training intervention maintain or wane over time? *Prevention Science*, *5*, 73–89.

Eddy, J. M. & Chamberlain, P. (2000). Family management and deviant peer association as mediators of the impact of treatment condition on youth antisocical behavior. *Journal of Consulting and Clinical Psychology*, *68*, 857–863.

Eyberg, M. E., Nelson, M. N. & Boggs, S. R. (2008). Evidence-based psychosoial treatments for children and adolescents with disruptive behavior. *Journal of Clinical Child & Adolescent Psychology*, *37*, 215–237.

Forgatch, M. S. & DeGarmo, D. S. (2002). Extending and testing the social interaction learning model with divorce samples. In J. B. Reid, G. R. Patterson & J. Snyder (Eds.), *Antisocial behavior in children and adolescents: A developmental analysis and model for intervention* (pp. 235–256). Washington, DC: American Psychological Association.

Forgatch, M. S., Knutson, N. & Mayne, T. (1992). Coder impressions of ODS lab tasks. Unpublished technical manual. Eugene, OR.

Forgatch, M. S. & Patterson, G. R. (2010). Parent Management Training – Oregon model: An intervention for antisocial behavior in children and adolescents. In J. R. Weisz & A. E. Kazdin (Eds.), *Evidence-based psychotherapies for children and adolescents* (2nd edn, pp. 159–178). New York: Guilford.

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*, *3*, 185–198.

Hagen, K. A., Ogden, T. & Bjørnebekk, G. (2011). Treatment outcomes and mediators of parent management training: A one-year follow-up of children with conduct problems. *Journal of Clinical Child and Adolescent Psychology*, *40*, 165–178.

Hintze, J. M. & Matthews, W. J. (2004). The generalizability of systematic direct observations across time and setting: A preliminary investigation of the psychometrics of behavioral observation. *School Psychology Review*, *33*, 258–270.

Lindahl, K. M. (2001). Methodological issues in family observational research. In P. K. Kerig & K. M. Lindahl (Eds.), *Family observational coding systems: Resources for systemic research* (pp. 23–32). London: Lawrence Erlbaum.

Lorber, M. F. (2006). Can minimally trained observers provide valid global ratings? *Journal of Family Psychology*, *20*, 335–338.

McDonald, R. P. (1999) *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

Mori, L. T. & Armendariz, G. M. (2001). Analogue assessment of child behavior problems. *Psychological Assessment*, *13*, 36–45.

Ogden, T., Forgatch, M. S., Askeland, E., Patterson, G. R. & Bullock, B. M. (2005). Implementation of parent management training at the national level: The case of Norway. *Journal of Social Work Practice*, *19*, 317–329.

Ogden, T. & Hagen, K. A. (2008). Treatment effectiveness of parent management training in Norway: A randomized controlled trial of children with conduct problems. *Journal of Consulting and Clinical Psychology*, *76*, 607–621.

Patterson, G. R. (1982). *Coercive family process*. Eugene, OR: Castalia Publishing Co.

Patterson, G. R. & Forgatch, M. S. (1995). Predicting future clinical adjustment from treatment outcome and process variables. *Psychological Assessment*, *7*, 275–285.

Prinz, R. J., Foster, S. L., Kent, R. N. & O'Leary, K. D. (1979). Multivariate assessment of conflict in distressed and nondistressed mother-adolescent dyads. *Journal of Applied Behavior Analysis*, *12*, 691–700.

Roberts, M. W. & Hope, D. A. (2001). Clinic observations of structured parent-child interaction designed to evaluate externalizing disorders. *Psychological Assessment*, *13*, 46–58.

Shavelson, R.J. & Webb, N.M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Snyder, J., Reid, J., Stoolmiller, M., Howe, G., Brown, H. & Dagne, G. (2006). The role of behavior observation in measurement systems for randomized prevention trials. *Prevention Science*, *7*, 43–56.

Storå, B., Hagtvet, K. A. & Heyerdahl, S. (2012). Reliability of observers' subjective impressions of families: A generalizability theory approach. *Psychotherapy Research*, *4*, 448–464.

Stubbs, J., Crosby, L., Forgatch, M. S. & Capaldi, D. M. (1998). *Family and peer process code: A synthesis of three Oregon Social Learning Center behavior codes* (Training manual.). Eugene, OR: Oregon Social Learning Center.

Waldinger, R. J., Schulz, M. S., Hauser, S. T., Allen, J. P. & Crowell, J. A. (2004). Reading others' emotions: The role of intuitive judgments in predicting marital satisfaction, quality, and stability. *Journal of Family Psychology*, *18*, 58–71.

Received 10 August 2012, accepted 24 August 2013