FAMILY PROCESS

Normativity and Desirability in Observational Assessments of Family Interaction

JAMES E. DEAL*

Issues of normativity (responding in a typical or average fashion) and desirability (the tendency for raters to endorse positive characteristics rather than neutral or more negative ones) are common in areas of the social sciences that frequently utilize profile correlations to measure dyadic similarity. They have implications for family scholars as well. In the present study, a pre-existing data set was used to make an initial, though limited, investigation into potential confounds of normativity and desirability for macrolevel observational assessments of family interaction. An empirical example is presented using q-sort ratings of family interaction, with variance in observational assessments decomposed into component parts. High levels of both normativity and desirability were found, indicating possible problems in terms of both reliability and validity of assessment. While the results provide an interesting beginning, they are limited due to the use of a q-sort methodology as well as an instrument with limited background and use. These limitations are discussed, as well as alternative interpretations for normativity and desirability and implications for future research.

Keywords: Normativity; Desirability; Observational Assessment; Family Process

Fam Proc x:1–12, 2018

The measurement of whole family functioning has often used an "objective outsider's" perspective (Olson, 1977), due at least partly to the belief that family members cannot accurately observe and report on ongoing interactions over time (Markman & Notarius, 1987; Wampler & Halverson, 1993). Kerig (2001) argued that observational methods are "uniquely suited" to studying family processes that involve relationships rather than individual characteristics (or individual perceptions of relationships). As Grotevant and Carlson (1989) note, a strength of observational assessment is the ability "…to describe the global relational structure or characteristics of the whole family…". Observational assessments, as compared to other types of data collection, "…require fewer assumptions and inferences, are less susceptible to various confounding influences, and reflect greater face validity and generalizability…" (Jacob, Tennenbaum, & Krahn, 1987).

Observational assessments of family and marital functioning typically fall into three types: microlevel coding, macrolevel coding, and mesolevel coding (Grotevant & Carlson, 1989; Kerig, 2001; Lindahl, 2001). Each of these addresses a different level of analysis. The focus of this paper is on macrolevel coding systems, which typically

*Department of Human Development and Family Science, North Dakota State University, Fargo, ND.

Correspondence concerning this article should be addressed to James E. Deal, Department of Human Development and Family Science, North Dakota State University, Department 2615, PO Box 6050, Fargo, ND 58108-6050. E-mail: jim.deal@ndsu.edu.

This research was supported by a grant from the National Institute of Mental Health, #MH39899, to Charles F. Halverson, Jr., and Karen Smith Wampler. The author is grateful to them for the use of these data.

Family Process, Vol. x, No. x, 2018 © 2018 Family Process Institute
doi: 10.1111/famp.12368

focus on more global phenomena (Lindahl, 2001). Systems of this type typically utilize a Likert scale format and require intensive documentation and training, as well as high levels of information processing and inference on the part of the rater (Grotevant & Carlson, 1989; Lindahl, 2001). With these high demands of raters come potential problems, however. Grotevant and Carlson (1989) identified six possible problems that could be attributed to the rater. First, raters may make errors of central tendency, or a tendency to assign values at the middle of the scale and to avoid values at the extremes. They may also be susceptible to a leniency/severity effect, a reluctance to assign values at one end of the scale but not the other. They may, for example, avoid rating families as highly negative but show no such reluctance in rating them as highly positive. Third, raters are susceptible to a contrast effect, a tendency to rate families/couples in the opposite direction from their own family. Fourth, raters are suspect to the logical error, in which the rater constructs logical relationships among scales so that families/couples receive similar ratings on them. For example, couples who are seen as high on positive affect are also seen as high on warmth, or cannot be seen as high on coercion. Raters also make proximity errors, in which they give similar ratings to two items that are placed together on the scale. Finally, they are susceptible to a halo effect, in which they construct a global impression of the family, then carry this across items and fail to discriminate among them. The results of both types of errors are reduced reliability and validity in the rating scales, due to restricted ranges and spurious correlations (Grotevant & Carlson, 1989).

While discussion of these potential problems is not found in the family science literature, it is found in a variety of other sources. Halo, leniency/severity, central tendency (also called restriction of range), and logical errors are found throughout the psychological literature discussing cognitive biases in the rating of individuals and objects, i.e., rater error (see, for example, Becker & Cardy, 1986; Berman & Kenny, 1976; Hoyt & Kerns, 1999; Jackson & Furnham, 2001; Kasten & Weintraub, 1999; Murphy & Blazer, 1989; Murphy, Jako, & Anhalt, 1993; Solomonson & Lance, 1997).

NORMATIVITY AND DESIRABILITY EFFECTS

There is another set of potential problems in the use of observational assessments, however, that has also not been widely recognized in the family science area, namely normativity and desirability effects. Normativity, also known as stereotype accuracy or typicality, is the tendency for individuals to evaluate themselves or others in a fashion that is culturally stereotypic (Kenny & Acitelli, 1994), that is similar to how the average person would respond (Wood & Furr, 2015). Normativity was first noted by Cronbach (1955) as one factor that can influence the validity of a dyadic index, and it has been most prominently discussed in literature that often utilizes intra-dyadic profile correlations to assess similarity (Furr, 2008; Kenny, Kashy, & Cook, 2006; Wood & Furr, 2016). Kenrick and Funder (1988) used it to reference the stability of many traits within a population, noting that its presence would mean that dyadic similarity on those traits would be a consequence of being a member of the population and not a characteristic of a specific dyad. Normativity can make members of a dyad appear to be similar not because they actually are, but because both rate themselves in a stereotypic fashion (Kenny et al., 2006). It captures "the degree to which a profile reflects an average profile" (p. 1270, Furr, 2008), or those characteristics that are common to most individuals (Wood & Furr, 2016). As Furr notes (2008), all profiles are likely to be characterized by normativity to some degree (see also Kenny & Acitelli, 1994; Kenny et al., 2006; Wood & Furr, 2016).

Desirability simply refers to the tendency for raters to endorse positive characteristics rather than neutral or more negative ones. As Wood and Furr note (2016), desirability and normativity tend to be highly correlated—i.e., items endorsed more on average tend to be those that are more desirable.

From an analytical perspective, when normativity and desirability are removed from individual profiles two things happen. First, estimates of similarity between individuals tend to be significantly lower, as normativity artificially increases the degree of similarity between individual profiles (Furr, 2008; Humbad, Donnellan, Iacono, McGue, & Burt, 2013; Kenny & Acitelli, 1994; Kenny et al., 2006; Wood & Furr, 2016). Second, correlations between similarity estimates and other variables tend to change, often becoming significantly lower or even nonsignificant (Deal, Halverson, & Wampler, 1999; Kenny et al., 2006).

There is a limited history of these effects in the family science literature. Deal, Halverson, and Wampler (1989), in discussing parental similarity on childrearing orientations, found that similarity among randomly paired parents was as high, on average, as that between actual parental dyads. In follow-up research Deal, Halverson, and Wampler (1999) applied a method developed by Kenny and Acitelli (1994) to decompose parental similarity into dyadic-level similarity and stereotype similarity, i.e., normativity. When Deal, Halverson et al. (1999) removed normativity from their parental similarity profiles, they found that correlations between parental similarity and other measures of parenting dropped significantly, a finding shared by others working with similarity indices in other areas (Kenny et al., 2006; Wood & Furr, 2016).

While this discussion has been conducted in the context of similarity scores, the issues raised by it are relevant beyond simply the appropriate derivation of such scores. Effects such as normativity and desirability are potentially present in any rating of self, other, or object, so they may be present in the observational assessment of family interaction as well. To the degree that normativity and desirability are present in observers' ratings, reliability and validity would be expected to suffer. To date, however, no empirical investigation of these potential effects has been conducted.

Interpretation of Effects

Given the literature on normativity and, to a lesser degree, desirability, a key question is how to interpret the two constructs. There would appear to be at least two ways to do so. The first, as discussed above, is simply to view it as a methodological issue, i.e., as rater error or bias, a tendency for raters to rate families in the same, normative way.

However, it is possible that this is more than simply rater error. As Wood and Furr (2016) note, while researchers in personality psychology are increasingly removing the normative profile from their participants' data there remains a group who do not. The latter base their choice on the belief that doing so removes something meaningful rather than simply an artifact of score construction. In the present context, this requires us to view normativity and desirability not as characterizing rater error, but as characterizing the family's behavior in the observational setting. Deal, Halverson et al. (1999) noted that high levels of normative similarity between parents could be viewed as a link between the parents and the standards found in the larger social environment. They cited Kenny and Acitelli's (1994) view of stereotype similarity not as a bias, but as a typical way of responding. The same could be said here, that families displaying a high degree of normativity are simply showing an understanding of the norms of family interaction expected in a public area, and are appropriately displaying those norms. This reflects the view that normativity is a desirable trait, connected to a variety of positive characteristics (Wood & Furr, 2016).

Purpose

There is a need, then, to investigate issues of normativity and desirability in both the observational assessment of family interaction, as well as in the nature of the interaction itself.

The purpose of this report was to begin to examine the presence of normativity and desirability in ratings provided by trained observers using a macrolevel, observational q-sort assessment of family interaction. Analyses will focus on variance in the Q-sort data accounted for by normativity and desirability, contrasting it with remaining variance, on relations between these components, and on stability of each over time. As noted previously, macrolevel coding systems in family science tend to utilize Likert-style ratings. For the present study, data were available from an observational study of family interaction that utilized a Q-sort instrument rather than the typical ratings scales. The Q-sort methodology is generally well known in the child development literature (see Block, 2008), but much less known in the family science area. In Touliatos, Perlmutter, and Straus's (2001) original cataloging of family assessment instruments, for example, only two utilized a Q-sort. The Q-sort is a unique methodology, and the results presented here are limited by its use. They are also limited by the use of this particular instrument, as its presence in the literature is extremely limited. This study should, then, be considered exploratory, with the results hopefully pointing to areas of potentially fruitful further investigation.

METHOD

Participants

Data used in this study were originally collected as part of a 5-year study of (initially) preschool children and their families. (Full details of the recruitment and data collection procedures can be found in Wampler, Halverson, & Deal, 1996.) Participating families all had at least one child between the ages of 3 and 6 (the target children), no children older than age 10, and no children from any previous marriage(s). When more than one child fit the target age range, selection was random. Written informed consent was obtained from all parents prior to the child's participation in the study. For each year of data collection, families participated in an evening session in an observational laboratory and completed an extensive questionnaire packet. Families were paid for their participation in all years.

One hundred thirty-six families participated in the first year of the study. Due to changes in the observational measures after the first year, analyses in this report are based on the ninety-five families who completed both the second and third years of the study. These families were largely white (89.2% of both husbands and wives) with US citizenship (96.8% of both husbands and wives), educated (all husbands had completed high school, with 29% having a college degree and 52.7% having a graduate degree; all wives had completed high school, with 29.1% having a college degree and 37.6% having a graduate degree), and employed (92.7% of husbands were employed full-time; 23.75 of wives were employed part-time and 30.1% were employed full-time). Mean age of husbands was 35.1 years (SD = 4.9 years), with mean age of wives 33.34 years (SD = 4.1 years). It was a first marriage for 90.3% of husbands and 84.9% of wives, and couples had been married an average of 9.7 years (SD = 3.4 years). All children were from the present marriage, and the target children enrolled in the study had a mean age of 65.47 months (SD = 11.31); 54.3% were male, with 45.7% female.

To examine possible differences between this group of families and the initial group who were recruited, t-test and chi square comparisons were made between the sample used in these analyses and the 42 families from year one who did not participate in years two and three. No significant differences were found on any variable, though several of

the chi square analyses could not be used due to too many empty cells (i.e., citizenship, race).

Measures

Family interaction

While an observational assessment of family interaction using macrolevel ratings scales was desired, one was not available to the author. The data that were available measured family interaction utilizing the Georgia Family Q-Sort (Wampler, Halverson, Moore, & Walters, 1989). This 43-item q-sort was used to code laboratory interaction in which the mother, father, and target child built a house together of plastic construction blocks. Blocks were constructed in such a way that certain small pieces had to be used to connect larger blocks. In the first half of the task, families were given a limited amount of time to build a house that matched a model house given to them. Families were told that only the child could place these smaller blocks, to insure the involvement of the child, and a limited amount of time was allocated to this task (7 minutes). These restrictions were placed in order to make the task moderately stressful for families. In the second half of the task, they were asked to build a house of their own design, with no restrictions on who could place particular blocks, and no time limit. The entire procedure typically took approximately 15 minutes, though parents were allowed to take as long as desired on the second task.

Sample q-sort items are: "distinct division of labor," "parents ignore child," and "tense about accomplishing task." Items were sorted into nine categories using a quasi-normal distribution that ranged from "least like the family" (three items) to "most like the family" (three items), with "neutral or not salient" as the midpoint (seven items) (see Block, 2008, for a discussion of the rationale for using a forced distribution). Q-sort coders were trained in the manner noted by Wampler, Moore, Watson, and Halverson (1989) in the Georgia Family Q-Sort training manual. Specifically, coders were given in-depth descriptions of individual items and the q-sort process, and were trained through an extensive process of joint sorts with a trainer as well as individual sorts of episodes that had been previously coded by the trainer. Discrepancies between the coder and the trainer were noted and discussed, in an effort to reach a common understanding of the codes and their application. When coders reached an acceptable level of reliability, they were allowed to code real data. Regular meetings to discuss issues related to coding were held with coders, with individual refresher training provided when necessary. Group-level refresher training was provided in three to four yearly sessions. All interaction episodes were independently coded by two coders, with reliabilities calculated between them. After the individual coding was completed, coders discussed item placement and arrived at a consensus sort for each family. Over the 4 years of the study, seven coders were used: three females (two of whom were graduate students, and one an undergraduate) and four males (three of whom were graduate students, with the remaining male a college graduate). In any given year, two to four of these coders were utilized. Wampler, Moore et al. (1989) have reported high levels of agreement between raters (mean Spearman-Brown for .77 for year 2 and .82 for year 3). Construct and criterion validity are all high (Wampler, Halverson et al., 1989).

As mentioned previously, Q-sort usage in family science is very limited, and usage of this particular Q-sort is even more limited. Results presented here must then be viewed very tentatively. They should not, in particular, be viewed as representative of macrolevel observational coding systems, but rather as a potentially interesting initial window into a question of interest.

Normativity

For each year of the study, an average, or normative, q-sort was created by calculating the means of each item. Items were then ranked from lowest to highest mean value, and the distributional norms of the instrument were applied—i.e., the three lowest items received a code of "1", the next four items received a code of "2", etc. The two normativity profiles were positively and significantly correlated ($r = .85$, $p < .000$), indicating that the construct is highly stable across 1 year. It should be noted that the normativity profile here is constructed from the same data as the individual rater profiles. While doing so is typical of these types of analyses, ensuring that the profile represents the individuals in the sample, it does mean that the two profiles are not completely independent. With the number of items in this sort, it is not statistically a problem; with smaller numbers of items, however, it could be (Eyvindson, Kangas, Hujala, & Leskinen, 2015).

Desirability

When the Family Q-Sort was originally created, an "ideal" q-sort was created as well. This was done by having ten faculty and graduate students in marriage and family therapy sort the q-sort deck in what they saw as an ideal family. Mean agreement across raters was .75 (range .67–.80), indicating a relatively high level of consensus. This was replicated with a second set of five expert raters, obtaining a mean level of agreement between raters of .73 (range .6–.86). This ideal sort was used as a profile of desirability. The desirability profile was significantly correlated with both normative profiles, $r = .78$ and $r = .76$, both $p < .000$.

Analyses

In order to determine the relative contributions of normativity and desirability to observational ratings of family interaction, a commonality analysis was conducted. Commonality analysis is used to decompose explained variance in a regression model when predictor variables are not orthogonal, and allows examination of explained variance that is unique to each predictor variable as well as that common to all possible combinations of predictor variables (Nimon & Gavrilova, 2010; Pedhazur, 1982; Zientek & Thompson, 2006).

To conduct this analysis, the data matrix for each set of Family Q-Sort profiles was transposed using SPSS. This results in a matrix in which variables are now rows and individual families are now columns. The normative and desirable profiles were then added to this matrix, as columns. Each individual family's q-sort profile was then regressed on the normative and desirable profiles. Multicollinearity tests from SPSS (tolerance and VIF) were both within acceptable ranges, indicating no multicollinearity present. The resulting R^2 indicates the amount of variance in the individual family profile accounted for by these two variables. In the SPSS regression procedure, the part correlations calculated between each predictor variable and the outcome variable are actually semipartial correlations; when squared, they represent the percent of variance in the outcome variable explained by that predictor variable alone, not shared with any other predictor variable (Pedhazur, 1982). Subtracting these squared semipartial correlations from the model R^2 gives the percent of variance explained by all predictor variables, in common (Pedhazur, 1982). Subtracting the R^2 from 1.0, of course, gives the amount of variance not explained by the predictor variables, and can be interpreted in this context as both remaining systematic variance and error variance. The focus of this report is on the variance explained by each of these components.

RESULTS

Descriptive statistics for all variance components are presented in Table 1. For the first year of data considered, normativity and desirability accounted for an average of 58% of the variance in the family q-sort profiles (*SD* = .22, range = .02–.87). Normativity uniquely accounted for an average of 19% of the total variance in the profiles (*SD* = .10, range = .00–.46), while desirability uniquely accounted for only 1% (*SD* = .02, range = .00–.09). Normativity and desirability commonly accounted for an average of 37% of the total profile variance (*SD* = .16, range = −.04 to .62). (Note: As Zientek and Thompson (2006) note, negative values are possible in commonality estimates. This is typically viewed as either sampling error or as suppression effects among independent variables. In these data, only three families had negative commonalities.) Finally, the average amount of variance in the profiles not attributable to either predictor variable, and thus viewed as remaining systematic variance and error variance combined, was 42% (*SD* = .22, range = 13–.98).

For the second year of data considered, normativity and desirability accounted for an average of 66% of the variance in the family q-sort profiles (*SD* = .16, range = .11–.88). Normativity uniquely accounted for an average of 24% of the total variance in the profiles (*SD* = .08, range = .03–.52), while desirability uniquely accounted for only 1% (*SD* = .01, range = .00–.08). Normativity and desirability commonly accounted for an average of 40% of the total profile variance (*SD* = .14, range = −.03 to .62) (in these data, only one family had negative commonalities). Finally, the average amount of variance in the profiles not attributable to either predictor variable, and thus viewed as remaining systematic variance and error variance combined, was 34% (*SD* = .16, range = 12–.89).

DISCUSSION

The present study is an initial effort to look at the potential effects of normativity and desirability in observational research on family interaction. A secondary dataset was used with a less than optimal observational measure, which limits the generalizability of the results. As such, this study should be viewed as a potential point of entry into this area, with a need for the results to be replicated with more typically used instruments.

In looking at the effects of normativity and desirability in profile similarity scores, Wood and Furr (2016) concluded that the effects were found virtually everywhere, across various scales as well as across different types of constructs. The results presented here

TABLE 1
Descriptive Statistics for All Variance Components

Component	Mean	*SD*	Min.	Max.
Year 2				
Average total variance accounted for	.58	.22	.02	.87
Variance unique to normativity	.19	.10	.00	.46
Variance unique to desirability	.01	.02	.00	.09
Variance common to both	.37	.16	−.02	.62
Variance remaining	.42	.22	.13	.98
Year 3				
Average total variance accounted for	.66	.16	.11	.88
Variance unique to normativity	.24	.08	.03	.52
Variance unique to desirability	.01	.01	.00	.08
Variance common to both	.40	.14	−.03	.62
Variance remaining	.34	.16	.12	.89

suggest that these effects extend beyond personality constructs and profile similarity scores to macrolevel, q-sort assessments of family interaction by trained observers. For both years of data presented here, normativity and desirability accounted for over half the variance in the observed q-sort data. Breaking that down in the commonality analysis, the majority of explained variance was shared by both normativity and desirability (37% in the year 2 data, 40% in the year 3), with another sizable portion attributed to normativity alone (19% and 24% respectively). Only desirability alone failed to account for significant portions of variance in the rated q-sort profiles (1% in both years). In this context, then, the effects of desirability are subsumed by those of normativity.

Interpretation of Effects

The question of how to interpret these results remains central. In most of the existing literature, normativity and desirability effects are viewed as characteristics of the raters and, subsequently, often as rater error. When applied to family interaction, however, it is clear that this is only one option. It is also entirely possible that normativity and desirability are characteristics of the family being observed, and of their behaviors in the observational setting.

If viewed simply as rater error, Furr (2008) and Wood and Furr (2016) offer some suggestions for dealing with normativity and desirability from this perspective. Most of these suggestions focus on the creation of profile similarity scores, primarily on calculating estimates for all components of similarity—that unique to the dyad as well as that due to normativity and desirability—and either analyzing all three components or controlling statistically for the effects of normativity and desirability. One suggestion, however, focuses on the rating procedure itself, and refers to creating instruments that control for normativity and desirability by removing items likely to be high on normativeness.

If normativity and desirability are characteristic of the family's behavior in the interactional setting, however, these suggestions are inappropriate. And if such behavior is viewed as either appropriate or as desirable, then attempting to influence the family's behavior to reduce it is also inappropriate. Instead, removing normativity from ratings of the family may result in emphasizing fundamentally different aspects of the family's interaction style. Wood and Furr (2016) note that, due to the high correlation between normativity and desirability, "...removing the normative profile increases the likelihood that the extreme elements of an individual's distinctive profile are more neutral or undesirable characteristics" (p. 5). Characterizations of interactive patterns would, then, be fundamentally altered with these normative elements removed, potentially providing a very different picture of the family under study.

Limitations of the Georgia Family Q-Sort

As previously discussed, the Georgia Family Q-Sort was used in this study because it was used in a pre-existing dataset available to the author. The general lack of use of this instrument in the field since its development certainly makes it a less than optimal choice. In addition, the general q-sort nature of the instrument must also be viewed as a limitation. The majority of macrolevel observational coding systems of family interaction use Likert-type scales. Given the q-methodology and the lack of use of the instrument, it is impossible to generalize these results to the broader set of macrolevel observational family assessment.

How Likert-type responses compare to q-sort responses is an empirical question, but it is one which has no literature in the family science area to bring to bear directly. There is a limited literature from other fields that examines the differences and similarities between q-sort and Likert formats when used to rate external events and stimuli. Much of

this literature focuses on comparing and contrasting results from q-factor analysis with the more traditional r-factor analysis. In these studies, results from the two methods are generally viewed as highly similar (Eyvindson et al., 2015; Havlikova, 2016; ten Klooster, Visser, & de Jong, 2008; Thompson et al., 2013), with certain caveats. Both Eyvindson et al. (2015) and Havlikova (2016), for example, note that the focus of the q-sort data on subjectivity can lead to different interpretations of very similar results, and Thompson et al. (2013) note that differences between the two can often be attributed to the typically smaller sample size in q-sort studies.

There are also limited direct comparisons between item ratings acquired from q-sorts and from Likert scales, with the two again being found to produce highly similar results (Eyvindson et al., 2015; Havlikova, 2016; ten Klooster et al., 2008; Thompson et al., 2013). ten Klooster et al. (2008), for example, found a correlation of .93 between q-sort and Likert ratings, noting that the method used "did not substantially affect the way respondents rated the 30 items overall" (p. 516). In both factor analysis results and item ratings, then, q-sort and Likert methodologies seem to provide highly comparable results when used to rate external events. Again, however, this literature is outside of the focus of the present paper, so the results presented here must be viewed as limited.

It is also important to recognize that the study used for these data utilized a consensus-based final assessment of the family. It is possible that this might have increased levels of normativity or desirability in these ratings. In addition, researchers utilizing observational assessments of family interaction often use either microlevel or mesolevel coding systems, rather than macrolevel, and the level of specificity (vs. abstractness) of descriptors used can vary both across and within types of assessments. The role that normativity and desirability may play in these different systems and at these different levels is currently unknown.

Impact of the Observational Setting and Task

The nature of the family interaction observed, specifically the setting, the task used to elicit family interaction, and the family members involved, must also be considered. As Jacob et al. (1987) have noted, the task that family members are asked to engage in imposes a structure on their behavior, both behaviorally and emotionally, and different tasks would be expected to impose different structures. Tasks also typically differ as to the degree of ecological and external validity represented (Lindahl, 2001). In terms of who is present, conceptualizations of "family" can range from dyads to triads to much larger groups, with coding and analytical complexities increasing with family size (Lindahl, 2001). In addition, different configurations of family groupings can lead to different behavioral levels and styles by family members, even when behavior is consistently coded toward the same individual—i.e., mother's behavior towards father in a dyad, versus mother's behavior towards father in a mother-father-child triad (Deal, Stanley Hagan et al., 1999). Finally, demand characteristics of the observational setting such as the type of room used and the setup of that room have an impact on family members' behaviors. It is possible that certain tasks, settings, and family groupings may be related to higher or lower levels of normativity or desirability in either observer ratings or family behaviors. As yet, these possibilities remain unexplored.

Limitations of the Sample

Finally, there are also a number of limitations present in the sample. This report relies heavily on white well-educated families. Research with a more diverse sample is needed to clarify how extensive these effects are. As noted, normativity may be a characteristic of either raters or of the families being rated. In either case, levels of normativity found may,

to some extent at least, be influenced by characteristics of the sample. What is normative in one sample might not be normative in another, and what is viewed by raters as normative in one sample may not be viewed the same way in others. How this would impact ratings or behaviors of the families in the sample is not clear, but is worthy of further exploration.

Implications for Researchers and Practitioners

With these points in mind, consideration of how to interpret current or previously published research that may be impacted by normativity and desirability is limited in the literature, and is all in the context of similarity indices. Certainly, when similarity indices are used, there is ample evidence that the levels of similarity indicated are likely to be inflated, as are relations with other variables. In addition, earlier research from both personality psychology and family science has consistently found that correlations between profile similarity scores and outcome variables typically drop once normativity is removed from the profile similarity scores. One reason for this, in the personality area at least, is the higher presence of normative items in the outcome measures (Wood & Furr, 2016). The same thing is true here; in both years, items such as "enjoy being together," "all cooperate in completing task," and "relaxed, comfortable with each other" were among the highest rated items in the normativity profile, with items such as "don't get along with each other" and "parents fight each other for control" among the lowest. Removing these items would certainly be expected to attenuate the correlations of any scale scores with other outcome variables. Although these findings are consistent, they are all based on the use of profile similarity scores. The impact on scale scores created from macrolevel observational ratings—or self-report ratings—is less clear, however, and there is no literature to guide researchers at this point.

It will also be critical for researchers and practitioners to remember that, while observational characteristics of family interaction appear to be biased in the direction of more normative and desirable behavior, the source of that bias cannot be definitively stated at this point. Determining the source will require a research design specifically created to explore three potential sources—rater, family, and setting—simultaneously. This design would require a large set of raters evaluating a large set of families in multiple settings, so that individual differences in levels of normativity and desirability attributable to raters, families, and the observational setting could be estimated and compared.

Along the same lines, another important line of future research would lie in understanding the common and unique characteristics of both components of variance, as well as investigating more potentially informative ways that they relate together beyond those presented here. It is also worth noting that while the majority of variance explained was, on average, shared between normativity and desirability, normativity continued to contribute significantly beyond the shared component while desirability did not. Further investigation to follow up this finding is also warranted.

SUMMARY

The results here, while limited by the use of an obscure instrument and a q-sort methodology, indicate that normativity and desirability may create problems for researchers and practitioners utilizing observational ratings of family interaction. Future research will be necessary to followup on the questions raised utilizing more normative and conventional assessments.

REFERENCES

Becker, B., & Cardy, R. (1986). Influence of halo error on appraisal effectiveness: A conceptual and empirical reconsideration. *Journal of Applied Psychology*, *71*, 662–671. https://doi.org/10.1037/0021-9010.71.4.662

Berman, J., & Kenny, D. (1976). Correlational bias in observer ratings. *Journal of Personality and Social Psychology*, *34*, 263–273. https://doi.org/10.1037/0022-3514.34.2.263

Block, J. (2008). *The Q-sort in character appraisal: Encoding subjective impressions of persons quantitatively*. Washington, DC: American Psychological Association. https://doi.org/10.1037/11748-000

Cronbach, L. (1955). Processes affecting scores on "understanding of others" and "assumed similarity". *Psychological Bulletin*, *52*, 177–193. https://doi.org/10.1037/h0044919

Deal, J. E., Halverson, C. F., & Wampler, K. S. (1989). Parental agreement on child-rearing orientations: Relations to parental, marital, family, and child characteristics. *Child Development*, *60*, 1025–1034. https://doi.org/10.2307/1130776

Deal, J. E., Halverson, C. F., & Wampler, K. S. (1999). Parental similarity on childrearing orientations: Effects of stereotype similarity. *Journal of Social and Personal Relationships*, *16*, 87–102. https://doi.org/10.1177/0265407599161005

Deal, J. E., Stanley Hagan, M., Bass, B., Hetherington, M., & Clingempeel, G. (1999). Marital interaction in dyadic and triadic contexts: Continuities and discontinuities. *Family Process*, *38*, 105–115. https://doi.org/10.1111/j.1545-5300.1999.00105.x

Eyvindson, K., Kangas, A., Hujala, T., & Leskinen, P. (2015). Likert versus Q-approaches in survey methodologies: Discrepancies in results with same respondents. *Quality & Quantity*, *49*(2), 509–522. https://doi.org/10.1007/s11135-014-0006-y

Furr, R. M. (2008). A framework for profile similarity: Integrating similarity, normativeness, and distinctiveness. *Journal of Personality*, *76*(5), 1267–1316. https://doi.org/10.1111/j.1467-6494.2008.00521.x

Grotevant, H., & Carlson, C. (1989). *Family assessment: A guide to methods and measures*. New York: Guilford.

Havlikova, M. (2016). Likert scale versus Q-table measures—A comparison of host community perceptions of a film festival. *Scandinavian Journal of Hospitality and Tourism*, *16*, 196–207. https://doi.org/10.1080/15022250.2015.1114901

Hoyt, W., & Kerns, M. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods*, *4*, 403–424. https://doi.org/10.1037/1082-989X.4.4.403

Jackson, C., & Furnham, A. (2001). Appraisal ratings, halo, and selection: A study using sales staff. *European Journal of Psychological Assessment*, *17*, 17–24. https://doi.org/10.1027//1015-5759.17.1.17

Jacob, T., Tennenbaum, D. L., & Krahn, G. (1987). Factors influencing the reliability and validity of observational data. In T. Jacob (Ed.), *Family Interaction and Psychopathology* (pp. 297–328). Boston, MA: Springer. https://doi.org/10.1007/978-1-4899-0840-7

Kasten, R., & Weintraub, Z. (1999). Rating errors and rating accuracy: A field experiment. *Human Performance*, *12*, 137–153. https://doi.org/10.1080/08959289909539864

Kenny, D., & Acitelli, L. (1994). Measuring similarity in couples. *Journal of Family Psychology*, *8*, 417–431. https://doi.org/10.1037/0893-3200.8.4.417

Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic data analysis*. New York: Guilford.

Kenrick, D., & Funder, D. (1988). Profiting from controversy: Lessons from the person-situation debate. *American Psychologist*, *43*, 23–34. https://doi.org/10.1037/0003-066X.43.1.23

Kerig, P. (2001). Introduction and overview: Conceptual issues in family observational research. In P. Kerig & K. Lindahl (Eds.), *Family observational coding systems: Resources for systemic research* (pp. 1–22). Mahwah, NJ: Erlbaum.

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

Markman, H. J., & Notarius, C. I. (1987). Coding marital and family interaction. In T. Jacob (Ed.), *Family interaction and psychopathology* (pp. 329–390). Boston, MA: Springer. https://doi.org/10.1007/978-1-4899-0840-7

Murphy, K., & Blazer, W. (1989). Rater errors and rating accuracy. *Journal of Applied Psychology*, *74*, 619–624. https://doi.org/10.1037/0021-9010.74.4.619

Murphy, K., Jako, R., & Anhalt, R. (1993). Nature and consequences of halo error: A critical analysis. *Journal of Applied Psychology*, *78*, 218–225. https://doi.org/10.1037/0021-9010.78.2.218

Nimon, K., & Gavrilova, M. (2010. March 2). *Commonality analysis: Demonstration of an SPSS solution for regression analysis*. Retrieved from http://hdl.handle.net/2142/15062

Olson, D. H. (1977). Insiders' and outsiders' views of relationships: Research studies. In G. Levinger & H. Rausch (Eds.), *Close relationships: Perspectives on the meaning of intimacy* (pp. 115–135). Amherst, MA: University of Massachusetts Press.

Pedhazur, E. J. (1982). *Multiple regression and behavioral science: Explanation and prediction*. New York: Holt, Rinehart, and Winston.

Solomonson, A., & Lance, C. (1997). Examination of the relationship between true halo and halo error in performance ratings. *Journal of Applied Psychology*, *82*, 665–674. https://doi.org/10.1037/0021-9010.82.5.665

ten Klooster, P., Visser, M., & de Jong, M. (2008). Comparing two image research instruments: The q-sort method versus the Likert attitude questionnaire. *Food Quality and Preference*, *19*, 511–518. https://doi.org/10.1016/j.foodqual.2008.02.007

Thompson, A. W., Dumyahn, S., Prokopy, L. S., Amberg, S., Baumgart-Getz, A., Jackson-Tyree, J. et al. (2013). Comparing random sample Q and R methods for understanding natural resource attitudes. *Field Methods*, *25*, 25–46. https://doi.org/10.1177/1525822X12453516

Touliatos J., Perlmutter B. F., & Straus M. A. (Eds.) (2001). *Handbook of family measurement techniques: Abstracts* (Vol. *1*). Thousand Oaks, CA: Sage.

Wampler, K., & Halverson, C. (1993) Quantitative measurement in family research. In P. G. Boss, W. J. Doherty, R. LaRossa, W. R. Schumm, & S. K. Steinmetz (Eds.), *Sourcebook of family theories and methods: A contextual approach* (pp. 181–194). New York: Plenum Press. https://doi.org/10.1007/978-0-387-85764-0

Wampler, K., Halverson, C., Moore, J., & Walters, L. (1989). The Georgia Family Q-Sort: A new observational measure of family functioning. *Family Process*, *28*, 223–228. https://doi.org/10.1111/j.1545-5300.1989.00223.x

Wampler, K., Moore, J., Watson, C., & Halverson, C. (1989). *Manual of the Georgia Family Q-Sort*. Unpublished manuscript.

Wampler, K. S., Halverson, C. F., & Deal, J. E. (1996). Risk and resiliency in nonclinical young children: The Georgia Longitudinal Study. In E. M. Hetherington & E. Blechman (Eds.), *Stress, coping and resiliency in children and the family* (pp. 135–154). Hillsdale NJ: Lawrence Erlbaum Associates.

Wood, D., & Furr, R. M. (2016). The correlates of similarity estimates are often misleadingly positive: The nature and scope of the problem, and some solutions. *Personality and Social Psychology Review*, *20*, 79–99. https://doi.org/10.1177/1088868315581119

Zientek, L., & Thompson, B. (2006). Commonality analysis: Partitioning variance to facilitate better understanding of data. *Journal of Early Intervention*, *28*(4), 299–307. https://doi.org/10.1177/105381510602800405