2005

Personality and mental health as predictors of rater bias in observational data

Mark Richard Becker
Iowa State University

Personality and mental health as predictors of rater bias in observational data

by

Mark Richard Becker

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Psychology (Counseling Psychology)

Program of Study Committee:
Carolyn E. Cutrona, Co-major Professor
Daniel W. Russell, Co-major Professor
Douglas L. Epperson
Janet N. Melby
David L. Vogel

Iowa State University

Ames, Iowa

2005

Graduate College
Iowa State University

This is to certify that the doctoral dissertation of

Mark Richard Becker

has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Committee Member

Signature was redacted for privacy.

Committee Member

Signature was redacted for privacy.

Committee Member

Signature was redacted for privacy.

Co-major Professor

Signature was redacted for privacy.

Co-major Professor

Signature was redacted for privacy.

For the Major Program

TABLE OF CONTENTS

INTRODUCTION

Observational coding systems have been an important part of research in the social sciences for decades (Heyman, 2001). Observational coding refers to a method of data collection in which trained raters (or observers) watch and record what people actually do, in either a natural or laboratory setting (Whitley, 1996). The researchers using these systems have noted several advantages that this method of data collection provides. Most notably, observational coding affords more objectivity than self- or other-report measures, and thus the ratings provided by observational coders are more accurate than the data obtained from other types of data collection techniques (King, 2001).

However, observational coding systems do not always provide the objectivity that researchers hope for. A growing body of research has demonstrated that observational coding systems can be subject to rater bias (e.g., Becker, 1999; van der Valk et al., 2001). Rater bias is the systematic introduction of variance into the data set by the observers themselves – this variance is not attributable to the participants being studied, but rather is a function of the rater (Hoyt & Kerns, 1999). Research has shown that this bias varies considerably from rater to rater, and therefore it is difficult to control for by simply adjusting the coding data.

It is clear that this sort of variance is highly damaging, skewing the data in directions that do not accurately capture the behaviors being studied (Petkova et al., 2000). Although the effects of rater bias are potentially very damaging, few studies exist to help us understand what factors influence the emergence and degree of rater bias in observational data. Therefore, researchers are left with little direction in the search for ways to select or train raters that will ultimately minimize the rater bias present in observational data.

In an effort to better understand the phenomenon of rater bias, in this research I sought to examine facets specific to the rater that could be predictive of rater bias. It is documented in the social science literature that personality and mental health can have a significant impact on the ways in which people perceive events around them, even within research settings (Ambady et al., 1995). I performed exploratory analyses to assess the degree to which a variety of personality and mental health variables impact rater bias in a large, longitudinal observational data set.

LITERATURE REVIEW

Common techniques used by psychologists to collect data from research participants include administration of self-report or other-report measures, and the use of observational coding systems (Elmes, Kantowitz, & Roediger, 1992). Self-report measures require research participants to answer questions about themselves, whereas other-report measures request information about the participant from sources close to him or her (e.g., a spouse or close friend). Observational coding systems also use others to collect data about the research participants. However, these observers are typically not acquainted with the participant, but rather are trained members of the research team. Each of these methods has its own unique advantages and disadvantages, and the modality chosen for a given study is influenced by many factors, some theoretical, some practical. Whatever methodology a researcher chooses, however, it is very important that he or she has carefully considered the utility of the methodology chosen in order to produce data that are accurate and trustworthy. As Greenberg (1995) states, "An experiment is only as good as the observations on which it is based" (p. 366).

Observational coding consumes a great deal of time and money when compared to other methods of data collection. This cost is thought to be outweighed by the benefits such systems provide. The ability of observational coders to provide unbiased data is the key to the utility of observational coding systems, and is thus the main impetus behind their selection by researchers (King, 2001).

Given tight research budgets and the need for timely information, researchers must be assured that the benefit derived from using observational data collection methods is justified. Specifically, we need to be able to identify ways to ensure accuracy of ratings, and to identify

which characteristics are indicative of an unbiased rater. However, variables that compromise this advantage, such as the phenomenon of rater bias, are still largely ignored by researchers using observational data (Hoyt, 2002).

Individual differences, such as personality, mental health, and life experiences have been demonstrated to affect most facets of our perceptions and behaviors (Funder, 1999). Although people like to believe that they are impartial observers of the world around them, it is rarely the case that we can see ourselves, others, or situations around us as clearly as we might like to believe. Indeed, the psychological and sociological literatures have cited numerous situations and conditions in which we do not observe as accurately as we think we do (Cook, 1989).

When examining our own abilities versus the abilities of others to be objective observers, researchers have noted that we easily identify bias in others around us, but rarely see it in ourselves. Furthermore, when we do see bias in ourselves we are likely to attribute our distortions to insight that we have that others do not (Pronin, Yin, & Ross, 2002).

Considering this human tendency toward biased observation, it seems logical that the factors that fuel this phenomenon could impact the ratings given by observational coders. The presence of bias in the observations made by trained coders has been documented (e.g., Becker, 1999; Hoyt, 2002), and it is not at all new for researchers to be concerned about the accuracy of the ratings on which they base their research. As Lippa and Dietz (2000) note, some limited research exists dating back to the 1950's that has focused on identifying those factors that are predictive of accurate observations by raters, and in identifying what makes a "good judge" (p. 25).

Although a small number of researchers are examining these issues, the research literature has largely ignored the problem of rater bias in the context of observational coding, and has not provided many theories to account for the bias seen in the observational literature (Hoyt, 2000). In addition, there has been very little research examining the connection between the personality and mental health of coders and their rating accuracy, and the literature that does exist has provided inconsistent findings regarding which factors affect observer ratings (Colvin & Bundick, 2001). Colvin and Bundick warn that the number of studies in this area is so small that any judgments based on the existing literature must be viewed as tentative.

Becker (1999) demonstrated that rater bias was present in a well-established observational coding system. Coders went through 180 hours of training in the use of this well-defined system. Despite this very high level of training (much greater than that provided in most settings) and the fact that a long-established and well-respected coding system was used, Becker (1999) found significant amounts of rater bias in the data. The explanations that would commonly be proposed to explain the bias (e.g., unproven system, lack of rater training) were not applicable in this case. In addition, not all coders exhibited bias on the same scales, which negates the argument that the system is written in a way that predisposes certain scales to bias.

These findings lead to the possibility that the observed bias is attributable to individual differences in the coders. In an effort to better understand what would account for the observed rater bias, individual differences across raters can be explored to see what factors might distinguish a "good" coder from a "bad" one. By better understanding the role

of personality and mental health in rater bias, we can provide guidance for those using observational data to ensure their data are accurate and worthwhile.

In this literature review, I will first provide an outline of different types of data collection methods and observational coding systems. After describing the basic underpinnings of these systems, I will discuss the general advantages and disadvantages inherent in such systems, with additional attention paid to the specific problem explored in this research (i.e., rater bias). I will discuss generalizability theory as a methodology by which rater bias can be measured. Finally, I will review the literature which suggests that individual differences such as coder personality or coder mental health might contribute to rater bias.

Observational Coding in Psychological Research

Behavioral coding systems originally grew from psychologists' suspicions that self- and other-report methods of data collection were failing to accurately capture important data (King, 2001). Observational coding as a method of data collection has been used in psychology since the 1920's, and its unique advantages have helped it to remain a popular choice among researchers (Heyman, 2001).

In its most general form, observational coding is a procedure by which trained observers (also called "coders" or "raters") watch and record what people actually do, in either a natural or laboratory setting (Whitley, 1996). More specifically, "The methodology of direct observational assessment is characterized by the use of coders, raters, or judges, who usually are not participants in the interpersonal system being studied and whose task is

to unitize and assign meaning to some aspect of [an interaction]" (Alexander, Newell, Robbins, & Turner, 1995, p. 355).

Observational researchers divide coding systems into two types: microlevel systems and macrolevel systems (Floyd, 1989; Markman et al., 1995). Microlevel coding systems (e.g., Hill & Stephany, 1990) assess behavior by attending to overt, observable actions on the part of the research participants (or targets); such behaviors might include, for example, head nods or eye shifts. In these systems, interactions between the participants are usually broken into discrete units, based upon predetermined time intervals or speaking turns (Alexander et al., 1995). The trained observer then classifies the types of interactions that occur within each unit. The individual observer does not assess the possible motivations behind the behaviors, but instead simply notes when these overt behaviors occur; the investigators are left to infer the meaning and motivation behind the observed behaviors.

Macrolevel coding systems (e.g., Melby et al., 1998) are more commonly used than are microlevel systems (Heyman, 2001). Like microlevel systems, macrolevel data are gathered by having the trained observer watch the overt behaviors exhibited by research participants. These systems go a step beyond microlevel coding systems, however, in that they allow inclusion of the rater's interpretation of the target's overt behaviors (e.g., vocal tone) as indications of the target's motivation. This interpretation is typically allowed only within a well-defined system with specific guidelines for the interpretation (King, 2001). Elliot (1991) pointed out that the four kinds of variables that are most often studied in macrolevel observational coding systems are: content (what the participant actually says to another interactor), action (what the participant does, such as making personal disclosures or supporting the other interactor), style/state (how the participant says or does things, such as

in a hostile or warm manner), and quality (how effectively the participant says or does things in the task, such as how effectively they support the other interactor).

Unless otherwise noted, when the term observational coding system is used in this dissertation, I am referring to macrolevel coding systems, as such a system was the focus of this research.

Advantages of observational coding systems

By using observational coding systems, researchers can overcome some of the potential difficulties that are common in self- or other-report methods of data collection. Supporters of observational coding systems claim that they promote consistent interpretation of rating items and reduce the influence of responses sets, as well as producing observations that are more objective than self-report measures (King, 2001).

One common difficulty that is faced by researchers using self- or other-report data is that research participants' interpretations of items and response options may be inconsistent from one participant to the next (Elmes, Kantowitz, & Roediger, 1992). One participant's interpretation of a response option such as "sometimes" can be very different from another's, and this difference calls into question the comparability of scores across participants. A benefit claimed for observational coding systems is that such systems provide consistently applied rating scales across all observers and subjects (Markman & Notarius, 1987). Trained observers have presumably been instructed in the use of the rating scale, and agree about what evidence should be used to arrive at a score. Thus, researchers using trained observers can have some confidence that rating scales are being applied consistently and correctly from one participant to the next (Melby & Lorenz, 1996).

Another possible advantage of observational coding is that trained raters are more likely to be objective in their evaluations than the target persons themselves or the acquaintances of the target persons (John & Robbins, 1994). Self- and other-report procedures rely on the recollections and opinions of individuals involved in the study. It is questionable whether participants can accurately assess themselves or people close to them without their personal opinions and feelings coloring those responses (Elmes et al., 1992).

Observational coding systems presumably reduce this problem, because the observers do not have the personal and emotional investment that research participants might. Therefore, the data they provide will not be subject to the distortions (either intentional or unintentional) that can be introduced by research participants. Indeed, research has demonstrated that there can be discrepancies between observer reports and self- or other-reports of the same interaction (e.g., Feinberg et al., 2001; Robinson & Price, 1980). This difference may be attributable to response biases or impression management on the part of participants (Floyd & Markman, 1983), distortions to which observational coding methods are presumably not subject. If the difference in score is indeed based on these self- and other-report biases, then observational coding should avoid the errors that will be present in the data based on self- and other-reports.

Disadvantages of observational coding systems

As previously noted, studies have found inconsistencies between self-reports of interactions and observer reports of behaviors in the same interactions (e.g., Margolin et al., 1985). Some argue that the difference between the results of the two data collection methodologies is indicative of the fact that observational coding has succeeded; that those

using this methodology have avoided the impression management bias sometimes found with self- and other-reports. However, other researchers disagree with this assertion, and have suggested that observer reports are less accurate than self- and other-reports (Funder & Colvin, 1997). One possible explanation for this lack of accuracy on the part of the coder lies in empathy theory, which holds that we form perceptions of others' behaviors and intentions by putting ourselves in their position and inferring what our own intentions would be. Those who are acquainted with observational targets have more knowledge of the person in question, and therefore can make more accurate inferences about their behaviors than those in a zero-acquaintance condition (i.e., an observational coder) who have no prior exposure to the target and can offer no special insight (Cook, 1979).

Other disadvantages that have been noted in observational coding systems include the high level of both financial and time resources required to procure observational data, the lack of behavioral consistency (i.e., are participants behaving in a manner consistent with their typical behavior, or has the introduction of the observer or camera caused them to alter their behavior), and the introduction of rater bias, a specific type of systematic error. Researchers must also contend with the fact that observational coding is often expensive. Researchers must train a group of observers to accurately code the behavior of the research participants and, if the observers are paid for their work, using observational coders can become a financial impossibility for many researchers.

Researchers must also contend with the large amount of time that is often spent in the coding process. When self-report measures are used, data can be gathered fairly quickly. Having observers rate an interaction between research participants (e.g., a parent and child discussing problems in the home, or a therapist and client participating in a counseling

session) will usually require hours of work, in addition to the time required to train the observers. This can result in a lengthy period of data collection, something that can be problematic for researchers (Hoyt & Kerns, 1999).

Judgments made about research participants based on limited samples of behavior may not be accurate (Kondo-Brown, 2002). Many personality measurements are made based upon the assumption that participant behavior will be at least somewhat consistent across occasions and situations (Moskowitz, 1986). Observational coding is based on the same assumption. Indeed, it is rare that researchers are interested solely in participant behavior during a specific interaction task. Rather, we assume that the behavior displayed by the research participant will be at least somewhat representative of his or her personality and behavior beyond the observational task. Moskowitz (1986) cites several studies in which personality traits observed on the part of study participants were not stable across different interactors or situations, thus illustrating the danger of generalizing traits of participants based upon their in-task behavior to other situations or occasions.

Finally, observational coding systems are susceptible to bias introduced by the raters (van der Valk et al., 2001). Few researchers would contend that humans are perfect observers. Indeed, as Hill et al. (1998) noted, "Raters do not function as neutral recorders of some physical reality. Rather, they are influenced in their ratings by a number of different factors" (p. 346). This rater-induced error will be the focus of the next section.

Bias in Observer Ratings

Rater bias is systematic error that is attributable to the rater. This can be due to either the raters' differential interpretation of the rating scale, or to their unique reactions to

particular targets (Hoyt & Kerns, 1999). In the following sections, I will first describe common approaches to the study of rater bias. I will then explore the implications of rater bias in psychological research. Finally, I will discuss factors that can impact the magnitude of rater bias, as well as characteristics of the rater that could influence the amount of rater bias introduced into the data.

Approaches to the study of rater bias

Traditional approaches. Two types of rater bias that are commonly studied are leniency bias and halo effects (Saal, Downey, & Lahey, 1980). Although leniency bias has been conceptualized differently by different investigators (Saal et al., 1980), the most useful definition from a psychometric standpoint pertains to raters' differential leniency – i.e., differences in the mean ratings assigned by raters across a large number of targets (Hoyt & Kerns, 1999). Thus, leniency bias, when present, reflects variance in ratings attributable to observers' differentially favorable evaluations of targets. For example, a rater who has a very positive outlook on life might consistently assign higher ratings to targets on scales such as warmth or positivity than are appropriate, based on the participant's actual behavior. By the same token, this coder might downplay negative behaviors exhibited. By definition, leniency errors are present across all targets rated by a given observer.

"Halo effects" refer to the tendency of the rater to attend to a global impression of each target, rather than to carefully distinguish among different levels of the target's behavior or performance (Borman, 1975). Halo effects are associated with rater-target interactions; that is, they are unique reactions that the observers have to specific targets. Research on halo effects has focused on the possibility that correlations between rating dimensions may be

inflated due to the impact of the rater's global impression (Feeley, 2002). For example, suppose a rater's job is to assess targets on levels of two unrelated constructs, like positive mood and mechanical aptitude. A rater might have a strong positive reaction to a target. As a consequence, the rater might then assign scores that are higher than appropriate across a range of positive attributes. Thus, the target might receive high scores on both positive mood and mechanical aptitude. Analysis of these data would indicate a correlation between these two attributes that is not accurate, but instead a reflection of the halo effect.

Generalizability approaches. Generalizability approaches provide a standard approach to assessing magnitude of rater bias in terms of the proportion of total variance in ratings that is attributable to raters rather than targets (Hoyt, 2002). Generalizability theory (GT; Cronbach, Gleser, Nanda, & Rajaratnam, 1972) is an extension of classical reliability theory that allows investigators to simultaneously examine the impact of several sources of measurement error (e.g., raters, items). GT analysis is similar to factorial ANOVA in that total variance is a function of main effects and interactions among the factors being examined. Instead of testing effects for statistical significance, GT analysis derives variance estimates for each main effect and interaction in the model, estimating the contribution of each source to the total variance in the ratings (Coates & Thoresen, 1978).

Consider the following example. Suppose four raters were rating 25 targets, using 10 items from a scale measuring hostility. GT analysis would provide variance estimates for three main effects (rater, target, and item), three two-way interactions (rater by target, rater by item, target by item), and one three-way interaction (rater by target by item; this is referred to as the residual, and is confounded with error). The target variance is the valid

variance in the model, as the target is the focus of the assessment. In addition to the variance components, GT also provides generalizability coefficients. Generalizability coefficients are derived by dividing the valid variance by the estimated total variance, and are interpreted as the proportion of variance that is due to actual differences among targets. The generalizability coefficient can be considered a measure of reliability; the higher the generalizability coefficient, the more reliable the rating scale.

When multiple observers rate multiple targets, rater variance (variance due to rater main effects) can be estimated using a generalizability analysis. Rater variance represents variance due to mean differences among raters, or individual differences in leniency. If multiple ratings of each target by each observer are collected (e.g., ratings on multiple items or multiple occasions) then rater-target interaction variance can be estimated separate from error. Rater variance and rater-target interaction variance are the most important sources of rater bias (Hoyt & Melby, 1999), and we therefore will explore these sources of variance further in the following section.

Impact of rater bias

Like other sources of measurement error, rater bias acts to reduce or attenuate observed correlations between variables, relative to the correlations that would be observed under conditions of error-free measurement (Petkova et al., 2000). For example, suppose one were interested in correlating self-report measures of drug use with observational data assessing quality of family interaction, to test the hypothesis that those persons coming from families exhibiting a high degree of desirable interaction (e.g., good communication, warmth) will be less likely to develop later drug addiction. The observed correlation will be

lower by a factor equal to the product of the square root of the reliability coefficients of the two measures, or $(\underline{r_{xx'}}\underline{r_{yy'}})^{1/2}$. So, for example, if the self-report drug use measure has a reliability coefficient of $\underline{r_{xx'}} = .70$ and the observational measure has a reliability coefficient of $\underline{r_{yy'}} = .60$, then the observed correlation would be attenuated by a factor of $[(.7)(.6)]^{1/2} = .65$. The effect of this attenuation will be the underestimation of the population correlation. If the true population correlation between the measures were $\underline{r_{xy}} = .3$, then the observed correlation (the attenuated correlation) would be (.3)(.65) = .20. The attenuation has resulted in an underestimation of the true population correlation.

This attenuation of observed correlations can impact other research considerations. Rater bias can, for example, reduce statistical power in much the same way that it attenuated the observed correlation in the previous example. When statistical power is calculated, researchers must take into consideration their potential sample size (*N*), the statistical significance level they desire to use (α), and the size of the population effect they are seeking (ρ) (Cohen, 1988). As we saw in the previous example, the population effect size will be attenuated by the unreliability of the instruments used. This attenuation must be considered when calculating statistical power. If a researcher decided to use an effect size of .3 (a moderate effect, according to Cohen) in his or her calculation to determine sample size needed for a certain significance level, he or she must remember to use the attenuated effect size. In this case, the attenuated effect size will be (.3)(.65) = .2, as it was in the previous example. If the researcher were to use .3 instead of .2, he or she would underestimate the number of participants needed to achieve the desired level of statistical power.

Factors affecting the magnitude of rater bias

The magnitude of rater bias present in observational data can be affected by a number of factors. These factors include dimensions of the rating scale itself as well as the relationship between observer and target.

Characteristics of the rating scale. In their examination of generalizability studies utilizing observer ratings, Hoyt and Kerns (1999) found that 37% of the variance in ratings could be attributed to rater bias. The level of bias in ratings was strongly moderated by several factors related to the rating scale. First, ratings of explicit attributes (as are often used in microlevel coding systems) were demonstrated to have negligible bias variance, whereas ratings requiring greater degrees of inference on the part of the rater (as are often used in the more common macrolevel coding systems) may contain substantial rater variance. Rater training also had a substantial moderating effect on rater bias for high-inference rating systems. Those laboratories in which investigators required more training (greater than five hours) of the observers demonstrated lesser amounts of rater bias than did laboratories requiring less training. However, studies have suggested that after an initial threshold of training time, increased rater training does not necessarily increase the accuracy of ratings (e.g, Bernieri & Gillis, 1995). Therefore, arguments often heard in the literature that rater bias is due to lack of training or an ill-conceived coding system are unlikely to completely explain rater bias.

Characteristics of the rater-target pair. The tendency to rate persons more favorably that we perceive to be similar to ourselves has been well documented in the psychological

literature (e.g., Baskett, 1973; Hill et al., 1988; Mahalik et al., 1993). If an observer feels that a target is similar to him or herself in some way, then he or she is likely to view that target in a positive light, and his or her ratings may be positively skewed. For example, in observations of parent-child interactions, an observer who is a parent may rate other parents more favorably, whereas non-parent observers may identify more strongly with the child.

The rater's subjective response to the target may also influence rater bias. If an observer has an overall positive reaction to an observational target, then it is likely that this rater will (consciously or unconsciously) inflate the target's scores on positive attributes or minimize negative attributes.

Individual Characteristics of the Rater Related to Bias

Research is very limited regarding individual characteristics of observers and the influence these characteristics have on rater bias, despite the potentially critical role these individual differences may play in rater bias. Research has shown that there can be bias present in data sets that is not consistent across raters (despite having gone through identical training and using the same coding system; Becker, 1999). When differential amounts of rater bias are seen across targets, independent of training received or system used, then one must look to the coder himself or herself in order to determine what would be causing this bias to be introduced into the data. Below I will outline some of the findings regarding individual differences among raters and the potential impact these differences have on rater bias; the reader is reminded, however, that this is an area that is largely unexplored, and therefore these findings should be considered preliminary.

Impact of intelligence

Raters' intelligence has been found to be positively correlated with accuracy of judgments when raters' scores were compared to "true" scores based on a consensus group of coders (Lippa & Dietz, 2000). Lippa and Dietz note that there are many possible explanations for this finding, including that higher intelligence can cause one to be more empathic (and increased empathy leads to more accurate ratings). Davis and Kraus (1997) performed a meta-analysis examining the relationship between several different intrapersonal measures and "empathic ability." Empathic ability was defined as the ability to judge others' emotional states, interpersonal relationships, and personality, and is identified as an important ability for observational coders to posses. Although they examined several predictors of empathic ability, they found that intelligence was most highly correlated with high scores on empathic ability. Other explanations for the link between intelligence and rating accuracy may be that higher intelligence can lead to being more observant; or that higher intelligence helps coders to learn and apply the coding system, thereby overriding some of the bias they might otherwise introduce.

Sex and gender roles

Sex has been explored as a factor impacting the accuracy of perceptions of others; there are many studies in the social science literature that have documented the impact that sex of either the perceiver or the perceived can have on perception (e.g., Swim, 1994). The research examining sex as it impacts observational coding is more limited. Although some studies have examined sex as it impacts the perception of others, these studies have often examined the sex of the target as the determinant of bias. Very few studies have examined

the sex of the coder as a predictor of bias. That is, while we have some information on how sex of the observational target impacts the ways in which observational coders perceive them, there is little work to guide us in the examination of how the sex of the observer might influence the coding he or she produces (Winquist, Mohr, & Kenny, 1998).

Within the limited range of findings in this area, it has been noted that women seem to be more accurate coders than men (Ambady, Hallahan, & Rosenthal, 1995). This is not unexpected, given that endorsement of traditionally feminine gender traits is correlated with greater judgmental accuracy, as well as self-report measures of empathy and interpersonal sensitivity (Cook, 1985). We assume that women exhibit more of these traditionally feminine gender traits than men, and therefore are more accurate coders.

Although it has been asserted that women are more accurate raters, women have also been found to rate targets more favorably across all five facets of the Big 5 personality traits when compared to male coders. According to one group of researchers, this finding is "consistent" (Winquist et al., 1998, p. 370). This finding was replicated when examining female coders' tendency to rate targets higher on facets that reflect positively on targets (e.g., warmth) than male coders do (Bettencourt, Dill, Greathouse, Charlton, & Mullholland, 1997). This is problematic, because this tendency to code targets in a more positive light than men do may cause women to miss important, if negative, information. The findings that women are more accurate, yet also artificially inflate their scores, appear contradictory.

This inconsistency is further complicated by studies that find no sex difference at all between men and women in terms of coding accuracy. For example, one group of researchers specifically looking for sex-linked bias found none (Lippa & Dietz, 2000). Becker (1999) also failed to find sex differences in coder bias. Further support for the

conclusion that there is not a sex-linked effect comes from the meta-analysis performed by Davis and Kraus (1997), which found that there was not a link between sex of coder and accuracy of observation. Once again, it is important to remember that the number of studies in this area is limited; therefore, this inconsistency is not unexpected, and one must take care when drawing conclusions from the literature.

The studies mentioned above have all examined biological sex as a determinant of coder accuracy. Masculine/feminine gender role endorsement exists independently from biological sex and from sexual orientation (Basow & Rubenfeld, 2003), and studies have typically not examined this characteristic as a predictor of coder accuracy. Clearly, this is an area that merits further research.

Impact of liking the target

Some have put forward the hypothesis that raters, "attempt to extract a general concept of the target as likeable or dislikeable" (Wyer, Lambert, Budesheim, & Gruenfeld, 1991, p. 98). Once coders form an initial impression of a target, they begin to interpret the target's behaviors through this filter. Similarly, Smith (1991) notes that raters tend to try to understand targets through the use of exemplars (i.e., a preset schema that is used to predict and explain the behavior of persons) and use these exemplars as shortcuts to explain the target's behavior.

For example, a rater who is made uncomfortable by teasing may explain the laughter of a target who is being teased as anxiety, even if the actual explanation is that the target finds the teasing funny. The coder might then go on to see other ambiguous behaviors by this target as indicative of anxiety once the rater has established in his or her own mind that

the target is anxious. Not all raters do this with the same frequency, however; some raters seem less prone to the use of exemplars (and therefore are capturing more true behaviors; Kahneman & Miller, 1986). Research has not examined what might explain why one rater uses these shortcuts whereas another rater does not. Furthermore, it has been demonstrated that the use of these shortcuts often occurs outside of observer consciousnesses, leading coders to believe they are basing their judgments on observed behavior when actually they are not (Kahneman & Miller, 1986).

Impact of intrapsychic factors

Popularized by Costa and McCrae (1986), the Five Factor model of personality (also referred to as the "Big 5") captures the facets that make up human personality under five broad categories: neuroticism, extraversion, openness, agreeableness, and conscientiousness (Aiken, 1999). The Big 5 model captures the rich diversity of human personality, yet provides a simple enough model to make conceptualization and measurement fairly straightforward. Because of this, the Big 5 is currently accepted as one of the standard models with which to conceptualize and examine global personality (De Raad, 2000).

The Big 5 is also considered a valid framework within the observational data literature. Lippa and Dietz (2000) note that research into observational coding issues has been assisted by, ". . . the emergence of the five-factor model of personality, which provides not only a framework for organizing the traits that judges are asked to perceive in others *but also for assessing personality traits of the judges*" (Lippa & Dietz, 2000, p. 26 [italics added]). The Big 5 has been used successfully in research in the areas of rater bias and observational data (e.g., Kenny, Albright, Malloy, & Kashy, 1994). However, the number of

studies in this area is very small. In addition, findings regarding the relation of Big 5 personality traits to the accuracy of observation have been inconsistent, as will be clear in the review of the literature that follows.

After conducting a meta-analysis of the judgmental accuracy literature, Davis and Krause (1997) concluded that psychologically well-functioning individuals (i.e., those with low neuroticism) were the best raters. Although this finding seems intuitively obvious, it should be noted that due to the lack of research in this area, their meta-analysis included many different kinds of judgmental accuracy studies, not just studies using raters observing an interaction task. For example, some studies used raters that were assessing individuals they knew well, as opposed to being in a zero-acquaintance situation such as is the case in most observational coding research. It cannot be assumed that the variables they extracted from the studies used in the meta-analysis were equivalent to one another due to the different types of studies included in the analysis.

Davis and Krause's (1997) assertion that well-functioning individuals are more accurate perceivers of others is contradicted by Ambady et al. (1995), who conducted a study in which coders were required to provide ratings of observational targets. They found that those who were the most accurate perceivers rated themselves as relatively low in expressiveness, social ability, and self-esteem. The authors concluded that individuals with social or personality impairment may be better judges of others than those who would be considered higher-functioning.

Other studies have found that there is "sketchy evidence" (Lippa & Dietz, 2000, p. 28) that the Big 5 variables of extraversion, agreeableness, and conscientiousness are linked to accuracy in observational ratings. In addition, Lippa and Dietz (2000) found that openness

was not predictive of accuracy as they had hypothesized it would be. John and Robins (1994) offered the provocative finding that a narcissistic view of self is associated with lower accuracy in assessing personality facets in others.

Additional individual difference variables of interest

Several of the studies examining personality and intrapsychic factors as they impact bias refer to psychological health or psychological well-being as being important dimensions when examining coder accuracy (e.g., Davis & Krause, 1997). However, none of these studies have used standardized measures of mental health to assess this relationship; instead, they infer mental health from other data or ask coders to rate themselves on various facets rather than use a standardized assessment tool (e.g., Ambady et al., 1995). Because it is widely known that mental health impacts the ways in which people perceive the world around them (Davison & Neale, 1994), this is an area that warrants further investigation.

Summary

Colvin and Bundick (2001) have summarized the current state of the research literature on personality facets, mental health, individual differences, and the accuracy of observer ratings. They state:

> "Tentative" describes the current state of research on judgmental accuracy. Inconsistent results . . . have led some researchers to call off the search for the good judge. Others are more optimistic, and see the problem as inherently important, and continue to search for the individual differences in judgmental accuracy. (p. 53)

It is clear that many factors impact our perceptions of others, and that this applies to observational coders as well. What is not clear when examining the literature is which specific factors influence coder accuracy, and the magnitude of their influence on rater bias.

Possible explanations for the lack of a consistent answer regarding how individual differences influence rater bias are many. First, there simply are not many studies that have examined how personality and other individual difference factors influence observational coders and, in turn, rater bias.

Second, there are many different ways to conceptualize and measure personality and mental health, as well as many different types of coding systems. Therefore, researchers are rarely looking at the same question, or using the same methodology, in their research. This "apples and oranges" approach greatly increases the chances that results will be inconsistent from one study to another. Replications of relationships are very rare, and therefore findings must be considered tentative pending replication. In addition, different studies offer conflicting explanations of what factors characterize a "good" coder, and therefore use similar terminology to examine different questions.

Another possible explanation for the lack of consistency in this area is that researchers have tended to focus on one narrow facet of personality or mental health when examining the impact individual differences have on the accuracy of ratings. Rather than conducting exploratory research to see what might be predictive of bias or measuring multiple dimensions of personality, researchers have sought to confirm or deny the impact of one specific facet. Because personality and mental health are multifaceted, this is an inefficient way to proceed.

I will conclude this literature review by noting that observational coding is a powerful tool for understanding human behavior, and one that has many benefits. I feel that this tool can be improved if we are better able to identify those factors that predict a "good" coder. By identifying those facets of the coder that predict bias, researchers will be able to be more selective in choosing coders, or to tailor which scales a specific coder is assigned to work on. These sorts of changes should lessen rater bias and increase the validity of the observational ratings.

Examining the impact that individual differences have on the expression of rater bias is an important addition to the literature in this area. The need for this research is clear. The proposed study will seek to further examine the link between personality, mental health, and rater bias.

METHODS

Participants

Observational coders

Forty-seven Family Interaction Analysts (observational coders) employed by the Institute for Social and Behavior Research (ISBR) at Iowa State University participated in this study. The observational coders were all Caucasian, and the majority of them were women (3 men were included in the sample). Their ages ranged from early 20's to mid-60's, with a mean age of approximately 34 years. All coders held a Bachelor's degree; five coders held a Master's degree. Approximately 55% were married, and 65% had children. Length of employment varied from 9 months to more than eight years, with a mean length of employment at time of data collection of approximately 2 years. They had all completed training on the Iowa Family Interaction Rating Scale (IFIRS; Melby et al., 1998).

Initial training of coders using the IFIRS consists of approximately 180 hours of instruction and practice coding under the supervision of coding unit administrative staff. This training includes didactic instruction in the use of the coding system, as well as viewing videotaped samples of the types of behaviors that will eventually be coded. Coders must pass a series of written and viewing tests at various points in their training in order to demonstrate proficiency with the IFIRS system before they begin coding tasks for the data sets. Coders are typically employed 20 hours per week, and have two training meetings each week with ISBR administrators to insure that they are coding in a manner that is consistent with the IFIRS Rating Scale. There is a system in place in which coders can ask ISBR administrators questions if they are unsure how to code a specific behavior.

Observational targets

All data for this study came from the Institute for Social and Behavioral Research (ISBR). This multidisciplinary research institute is staffed by professionals from a variety of social science and statistical disciplines, and is affiliated with Iowa State University. ISBR primarily focuses its efforts on studying the family, and makes extensive use of observational methods of data collection. Data for this study were obtained from the scores assigned to family interactions by coders at ISBR.

The majority of studies conducted at ISBR are longitudinal in nature, and involve very large samples; a single subset of data for one project might involve hundreds of families, each completing three or four separate interaction tasks. Therefore, it would be highly impractical to have all observers rate each family in a fully crossed design; such a design would consume an inordinate amount of resources. Instead, scores for each family are typically based on the ratings of a single observer. When only a single observer rates each target, nothing can be learned regarding interrater reliability (or generalizability). For this reason a subset (approximately 25%) of families are rated by two observers, to assess coder agreement.

The data sets used for this study were taken from two separate longitudinal studies conducted at ISBR. The focus of this research was the observational coders themselves; the observational targets were not variables of interest in this project. Therefore, these data sets were selected because the coding data were produced by observational coders in the study, and not because of any specific characteristics of the studies from which the observational targets were taken.

The first data set used is from the Iowa Youth and Families Project (IYFP). IYFP was designed to assess the impact of the farm crisis of the early 1980's on family interaction and youth outcomes. Data were taken from IYFP waves A through D, using tasks one through four. The dyads that make up these data involved either two adults in a discussion task, a parent and child in a discussion task, or a parent and child in a problem solving task. The discussion tasks used in this research were part of a more extensive data-collection interview that was conducted with each family. Two hundred and thirty seven families were included from the IYFP data, each of which had three interactors, for a total of 711 observational targets rated by the observational coders.

The second data set used for this study is from the Prosper Project. This project was designed to assess the effectiveness of intervention strategies to reduce drug use and other problem behaviors in adolescents. Data were taken from the Prosper pre-test, using tasks one and two. One hundred and fourteen families were included from the Prosper Project, each consisting of three family members, for a total of 342 observational targets.

In order to obtain the videotaped interactions, a trained interviewer travels to the participants' home to conduct the observational task. The interviewer sets up video recording equipment in a setting conducive for interactions among the observational targets, such as a kitchen table or a living room couch. Once the equipment is running and the participants are in place and have their instructions, the interviewer leaves the room.

Research participants are given a stack of stimulus cards that contain topics for discussion. The participants read the discussion topics, and to follow the instructions on the cards. The stimulus cards contain items such as, "What sorts of things do we enjoy doing together," and, "What does mom do when I do something she doesn't like." Questions are

selected in order to elicit both positive and negative interactions. The interviewer returns after a predetermined length of time, usually 15 - 30 minutes.

Measures

IFIRS Rating Scale

Coders in this study used the Iowa Family Interaction Rating Scales (IFIRS; Melby et al., 1998) to assign scores based on their global impressions of the interactor. Within this system, the coders note the overt behaviors that occur in the interaction task, and they work to interpret and attach meaning to these behaviors. It is important to note that these assigned meanings are based upon pre-determined and highly detailed definitions for each scale.

Consider the following example of how these coding scales are operationalized. The IFIRS defines anxiety as, "The extent to which the focal's verbal and nonverbal behavior communicates emotional distress that is conveyed as anxiety, nervousness, fear, tension, stress, worry, concern, and embarrassment. Person may appear tense, fearful, uncomfortable, and/or self-conscious. Attend carefully to nonverbal behaviors in scoring Anxiety" (Melby et al., 1998). The coding manual provides several pages of examples of words and behaviors observational targets might use that convey anxiety. The manual also outlines many auditory indications of anxiety.

The IFIRS coding system is comprised of 67 scales, broken into four general categories. The *Individual Characteristic Scales* assess interactors on dimensions independent of their interaction partner. Individual scales include scales such as Anxiety and Sadness. Both children and parents are rated on these dimensions. The *Dyadic Interaction and Relationship Scales* assess interactors on dimensions related to their interaction with

their task partner. Dyadic scales include scales such as Hostility and Assertiveness. Both participants are rated on these dimensions. The *Parenting Scales* assess interactors on their parenting skills and style. Parenting scales include scales such as Parental Influence and Positive Reinforcement. Only parents are rated on these dimensions. Finally, the *Individual and Group Problem Solving Scales* assess interactors on their problem solving skills and style. Problem solving scales include scales such as Effective Process and Solution Quality. Both participants are rated on these dimensions.

NEO Personality Inventory, Revised

All participating coders completed the NEO Personality Inventory, Revised (NEO-PI-R; Costa & McCrae, 1992). The NEO-PI-R consists of 240 items, and requires 35-45 minutes to administer. Participants respond to a series of trigger statements on 5-point Likert scales ranging from "strongly agree" to "strongly disagree." Trigger statements include items such as, "I have trouble making myself do what I should do," and, "I believe that laws and social policies should change to reflect the needs of a changing world." Results from the NEO-PI-R provide an assessment of the participant's personality within the Five Factor Model ("Big 5"); the factors assessed are neuroticism, extraversion, openness, agreeableness, and conscientiousness.

These scales are conceptualized as follows. Neuroticism is indicative of whether individuals are anxious, depressed, angry, emotional, and insecure. Extroversion reflects being sociable, gregarious, assertive, and active. Openness to experience is indicative of people who are imaginative, cultured, curious, original, and artistically sensitive. Agreeableness reflects being courteous, flexible, trusting, good-natured, and tolerant.

Finally, conscientiousness is indicative of those who are dependable, careful, thorough, responsible, and hard-working (Murphy, 1998). In addition to measuring the Big 5, the NEO-PI-R measures six specific facets within each of the five general domains, allowing more detailed analysis of personality than most other measures of the Big 5.

As noted in the literature review, the Big 5 model has been established as a useful and standardized way in which to assess personality. The NEO-PI-R allows "concise summary of an individual's emotional, interpersonal, experiential, attitudinal, and traditional styles" (Groth-Marnat, 1997). It has also been noted that the five factors appear very robust, and have been replicated in a number of studies using different methods of measurement. (Murphy, 1998).

Symptom Checklist 90, Revised

All raters completed the Symptom Checklist 90, Revised (SCL-90-R; Derogitis, 1997). This instrument contains 90 items, and takes approximately 15 minutes to complete. Participants are asked to indicate the degree to which a series of common symptoms are distressing to them. These responses are provided on 5-point Likert scales ranging from "not at all" to "extremely." Examples of the included items are, "feeling no interest in things," and, "feeling inferior to others."

The SCL-90-R screens for nine broad symptoms of psychopathology: somatization (distress arising from perceptions of bodily dysfunction), obsessive-compulsive (symptoms consistent with the psychological disorder of the same name), interpersonal sensitivity (feelings of inadequacy and inferiority), depression (symptoms consistent with the psychological disorder of the same name), anxiety (symptoms of worry, stress, and

apprehension), hostility (symptoms consistent with the negative affective state of anger), phobic anxiety (symptoms consistent with persistent, irrational fear states), paranoid ideation (symptoms consistent with a disordered style of thinking reflecting paranoia), and psychoticism (symptoms consistent with schizoid features, as well as first-order symptoms of schizophrenia).

The instrument also provides a global symptom index (designed to measure overall mental health), a global distress scale (designed to assess overall intensity of symptoms), and a global scale of positive symptomology (designed to assess the total number of active symptoms). The test is considered highly reliable and valid, and shows high correlations with MMPI items measuring pathology (Derogatis, 1997).

The SCL-90-R has been normed on both clinical and non-clinical samples, and standardized T-scores can be computed for members of either type of sample. This gives it an advantage over those measures of mental health that are normed against a clinical population only – when using the SCL-90-R a person does not need to meet the diagnostic threshold for mental illness in order for his or her maladaptive styles and behaviors to be measured, and for the impact of these styles on his or her functioning to be examined (Hersen & Turner, 1994).

Selection of Scales

IFIRS rating scales

While each observational target is typically coded on all scales that are applicable to the specific type of task they are participating in, researchers using ISBR coding data rarely use all of the data that is collected; it is neither efficient nor necessary to examine the vast

amount of data that is produced. Instead, researchers typically select scales that have been demonstrated to capture the dimensions in which they are interested. Therefore, I chose to examine only selected scales from the data sets, rather than all the scales that were present in the data.

The aggregation of the data sets eliminated the possibility of examining Parenting or Problem Solving Scales, because these scales are specific to a particular type of task and therefore are not represented for all tasks in the aggregate data set. Having narrowed the options to Individual and Dyadic scales (which are scored for all interactors on all tasks, and therefore were represented in each data set), I chose to use eight scales from the IFIRS system for analysis in this project: Hostility (HS), Angry Coercion (AC), Warmth/Support (WM), Assertiveness (AR), Listener Responsiveness (LR), Communication (CO), Prosocial (PR), and Antisocial (AN).

The eight scales used in these analyses were selected because they are most often used by researchers analyzing IFIRS data. An analysis of all studies published using IFIRS coded data indicates that these scales have been used in approximately 40% of studies, whereas the other scales are used significantly less frequently. Indeed, some of the scales that are coded have never been used in a published study (Frank & Anderson, 2004). By focusing on those scales that are most likely to be used by researchers at ISBR, I was able to increase the real-world applicability of the study while also working with a manageable amount of data.

The Dyadic Interaction Scales that were chosen are designed to assess interactors on dimensions related to their interaction with their task partner. Both children and adults are

rated on these dimensions. Brief definitions of the Dyadic Scales are listed below and are taken from the IFIRS coding manual (Melby et al., 1998).

- Hostility (HS): The extent to which hostile, angry, critical, disapproving, rejecting, or contemptuous behavior is directed toward another interactor's behavior (actions), appearance, or personal characteristics.
- Angry Coercion (AC): Control attempts that include hostile, threatening, contemptuous, or blaming behavior.
- Warmth/Support (WM): Expressions of interest, care, concern, support, encouragement, or responsiveness toward another interactor.
- Assertiveness (AR): The focal's ability, when speaking, to express self through clear, appropriate neutral and/or positive avenues using an open, straightforward, self-confident, non-threatening, and non-defensive style.
- Listener Responsiveness (LR): The focal's nonverbal and verbal responsiveness as a listener to the verbalizations or actions of the other interactor through behaviors that validate and indicate attentiveness to the speaker.
- Communication (CO): The speaker's ability to neutrally or positively express his/her own point of view, needs, wants, etc., in a clear, appropriate, and reasonable manner, and to demonstrate consideration of the other interactor's point of view. The good communicator promotes rather than inhibits exchange of information.
- Prosocial (PR): Demonstrations of helpfulness, sensitivity toward others, cooperation, sympathy, and respectfulness toward others in an age appropriate manner. Reflects a level of maturity appropriate to one's age.

- Antisocial (AN): Demonstrations of self-centered, egocentric, acting out, or out-of-control behaviors that show defiance, active resistance, insensitivity toward others, or lack of constraint; reflects immaturity and age-inappropriate behaviors.

NEO-PI-R scales

Because the facet of interest in this research is the effect of personality on rater bias, all five of the general domains on the NEO-PI-R were used (i.e., Neuroticism, Extraversion, Openness, Agreeableness, Conscientiousness). The subscales within each of the five domains were not examined, as this level of detail was not of interest in this study.

SCL-90-R scales

Three scales from the SCL-90-R were selected for inclusion in this analysis: obsessive-compulsive behavior (OCD), the Positive Symptom Distress Index (PSDI), and the Global Symptom Index (GSI). Because of the exploratory nature of this study, I was more interested in participants' global mental health (as measured by the PSDI and GSI). Therefore, these scales were selected for analysis. The obsessive-compulsive scale was included due to the unique nature of the coding tasks. That is, given that this position requires close observation and attending to small details, it followed that OCD might have an impact on coder behavior.

Design

Each family interaction was rated by two observers. The observers were randomly assigned to the families they rated, and therefore the pairings of observers were also random.

Observers did not know which of the tasks they were coding would be rated by another observer (and therefore used to assess reliability or rater bias), although they were aware that a percentage of their work would be selected for reliability analyses. Observers do not appear in the data set an equal number of times (as they would in a balanced incomplete block design; Fleiss, 1981) for to a variety of reasons (e.g., coders were employed differing numbers of hours per week, had differing tenure of employment).

Generalizability analyses often employ a fully crossed design, as this allows the most flexibility in the methods available to the researcher to estimate rater bias (Brennan, 1992). However, there are only one or two observational tasks per year that are coded by all raters at ISBR, and these data are typically not archived for future research. Therefore, I chose to use an incomplete block design in which each dyad was rated by two randomly selected raters, rather than all raters. This provided the advantage of being able to include many more targets and raters than if a fully-crossed design had been used.

Procedures

Observational coding

All observational tasks included were coded from videotapes; coders did not rate research participants as they interacted. This method ensures that the researchers have a record of the observations to which they can refer at any time in order to conduct further evaluations.

As noted earlier, approximately 25% of the tasks at ISBR are rated by two observers, in order to provide ISBR researchers data with which to check interrater reliability of the observational coding performed by the raters. Raters do not know which of their assigned

tasks are designated as Reliability Tasks. This helps to ensure that the raters are not giving special attention to the tasks that are bound for reliability; instead, researchers at ISBR can have some degree of certainty that the scores they obtain on a reliability task are a reasonable example of what is being submitted to the general data set.

Amount of time to code each interaction varies, based upon several factors such as the type of task coded and the complexity of the task. However, coders do have a maximum amount of time they are allowed to work on each interaction.

Assessment of coders

The NEO-PI-R and SCL-90-R data were collected from two different waves of coders. The first wave of data was collected in spring of 1995, whereas the second wave of data was collected in spring of 2004. This resulted in two different cohorts of coders, the first consisting of 28 coders, the second consisting of 22 coders. Three coders who completed assessment inventories were excluded from the study because scores on their inventories were not valid (items were left blank). Both sets of coders received equivalent training, and coded equivalent tasks.

Both assessments were administered during the coders' weekly training meeting, using a group testing format. Coders were not allowed to speak to one another during the assessment administration. Coders had the option of completing the assessments in a room with individual cubicles if they felt they needed more privacy. One coder chose to do this. Although both assessments had time limits, none of the coders exceeded these limits. Coders were able to ask the administrator questions about the instruments as they completed them.

Coders were only identified by a three digit number of their own choosing – no names were included on any assessment instruments, and this three-digit number was not included on the informed consent forms. Coders put their personally chosen three digit identification number and their ISBR coder identification number on a separate slip of paper; the coder ID number was then replaced in the data set with the three digit number the coders had chosen. This system ensured that no personally identifying information could be connected to the coders. In addition, it was stressed to coders that the Director of the Coding Unit, Lead Coders (administrators who manage specific coding groups), and the Director of ISBR would not have access to data linking their identities to their responses; this was done in order to help establish confidentiality and encourage honest responses.

All coders signed informed consent forms, and were provided debriefing information at the end of the test administration; these documents are presented in Appendices B and C. All phases of the data collection procedure were approved by the Iowa State University Institutional Review Board and Department of Psychology Human Subjects Review Committee (IRB # 03-218).

RESULTS

Preliminary Analyses

Several preliminary analyses were conducted in order to better understand the data, and to allow decisions to be made concerning the best ways in which the data should be analyzed. All preliminary analyses were performed using SPSS version 12.0 for Windows. Table 1 below provides means, standard deviations, and ranges for participants' scores on the predictor variables. All scores are represented as T-scores (i.e., M = 50, SD = 10) derived from comparison of raw scores to non-clinical normative samples.

Table 1. Descriptive statistics for predictor variables

Predictor	Minimum	Maximum	Mean	Std. Deviation
Neuroticism (N)	34	80	45.25	10.43
Extraversion (E)	33	72	51.15	9.44
Openness (O)	26	80	58.21	12.29
Agreeableness (A)	26	73	53.08	9.15
Conscientiousness (C)	20	69	49.09	9.44
Obsessive-Compulsive (OCD)	33	77	55.17	8.04
General Symptom (GSI)	35	80	53.96	7.93
Symptom Intensity (PSDI)	37	70	50.02	6.92
Negative Mental Health	37	76	53.76	6.91

Although the average scores center around a mean of 50 (as would be expected in a normal population), the rather large standard deviations indicate that there was considerable

variability in the data on these measures. Therefore, different personality types and levels of mental health were represented by the coders who participated in the study.

Equivalency of coding groups

As noted previously, the two groups of coders from whom data were collected received equivalent training and coded equivalent tasks; therefore, it was assumed that the groups would be similar to one another. However, since data were collected at different points in time, it was important to ensure that the groups were equivalent before they were combined and examined in aggregate. In order to verify that the two groups were not significantly different from one another on the predictor variables, an independent sample T-test was performed, assuming equality of variances. The results are presented in Table 2 below.

Table 2. Independent samples t-test for predictor variables

Predictor	*t*	df	*p*
Neuroticism (N)	-1.1842	45	.072
Extraversion (E)	.171	45	.865
Openness (O)	-1.270	45	.865
Agreeableness (A)	-.708	45	.211
Conscientiousness (C)	.891	45	.377
Obsessive-Compulsive (OCD)	.160	45	.874
General Symptom (GSI)	-.359	45	.722
Symptom Intensity (PSDI)	-.473	45	.639
Negative Mental Health	-.946	45	.349

As the table indicates, the two groups were not significantly different from one another on any of the measures. Therefore, the two groups of coders were combined together in subsequent analyses.

Effect of task

As noted previously, these data were derived from coding three types of interactional tasks: an adult romantic partner discussion task, a parent-child discussion task, and a parent-child problem solving task. These different types of tasks require targets to discuss different topics for different amounts of time.

Regardless of task type, all interactions were coded on the Individual and Dyadic scales, using the same IFIRS coding criteria (i.e., the coding system was not different depending on type of task). Effect of task type was examined to ensure that type of task did not impact rater bias scores across scales included in these analyses; task type proved not to impact rater bias on any scales included in these analyses. Because task did not have a significant impact on the degree of rater bias present in the data, the data were combined across the three types of tasks and examined in aggregate.

Correlations among the predictor variables

A correlation matrix was constructed to evaluate whether or not the predictor variables were independent of one another. Results are presented in Table 3 on the following page.

Table 3. Correlations among predictor variables

Predictor	N	E	O	A	C	OCD	GSI	PSDI
N	1 - 47	-.150 .315 47	.079 .600 47	-.325 .026 * 47	-.436 .002 * 47	.502 .000 * 47	.507 .000 * 47	.451 .001 * 47
E	-.150 .315 47	1 - 47	.302 .039 * 47	-.137 .360 47	-.029 .845 47	-.108 .470 47	-.172 .249 47	.057 .704 47
O	.079 .600 47	.302 .039 * 47	1 - 47	.202 .173 47	-.102 .497 47	.005 .973 47	-.105 .481 47	-.031 .838 47
A	-.325 .026 * 47	-.137 .360 47	.202 .173 47	1 - 47	.440 .002 * 47	-.149 .317 47	-.285 .052 47	-.139 .350 47
C	-.436 .002 * 47	-.029 .845 47	-.102 .497 47	.440 .002 * 47	1 - 47	-.121 .417 47	-.097 .518 47	-.048 .751 47
OCD	.502 .000 * 47	-.108 .470 47	.005 .973 47	-.149 .317 47	-.121 .417 47	1 - 47	.803 .000 * 47	.624 .000 * 47
GSI	.507 .000 * 47	-.172 .249 47	-.105 .481 47	-.285 .052 47	-.097 .518 47	.803 .000 * 47	1 - 47	.770 .000 * 47
PSDI	.451 .001 * 47	.057 .704 47	-.031 .838 47	-.139 .350 47	-.048 .751 47	.624 .000 * 47	.770 .000 * 47	1 - 47

Key: N = Neuroticism; E = Extraversion; O = Openness; A = Agreeableness; C = Conscientiousness; OCD = Obsessive-Compulsive; GSI = General Symptom; PSDI = Symptom Intensity
Note: * indicates significance

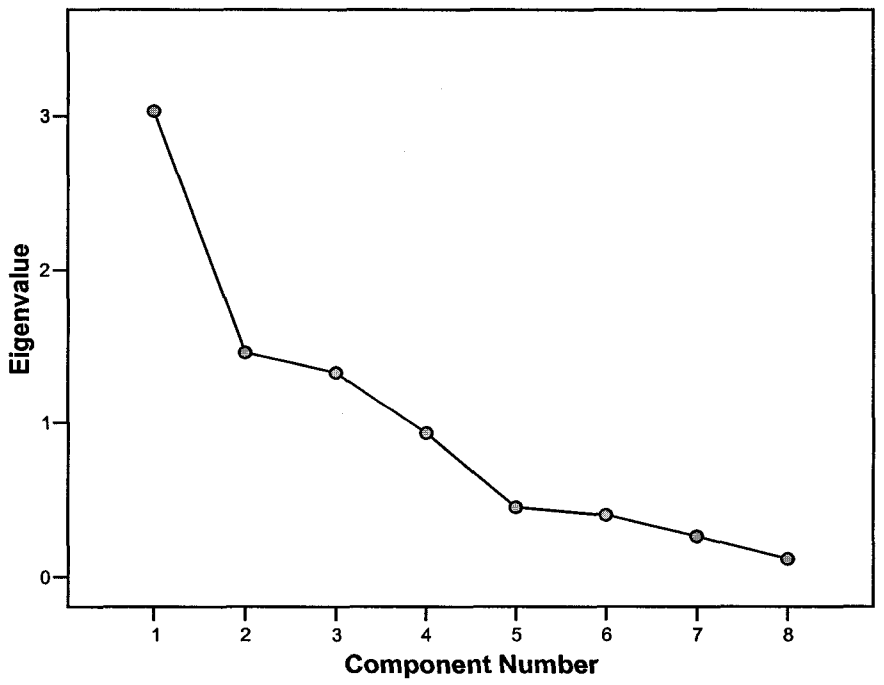
The correlation matrix indicates that there was a small, significant positive correlation (.3) between Extraversion and Openness. It also indicates that there was a moderate, significant positive correlation (.4) between Agreeableness and Conscientiousness. These correlations were not unexpected.

The matrix also reveals that there were several scales that were correlated with Neuroticism. Agreeableness showed a small, significant negative correlation with

Neuroticism (-.325), and there was a moderate, significant negative correlation with Conscientiousness (-.436). Neuroticism showed a moderate positive correlation with the three other measures of mental health from the SCL-90-R (r = .45 to .50). All three scales from the SCL-90-R were also highly correlated with one another (r = .62 to .80). These results indicate that coders scoring high on Neuroticism were also experiencing current, active symptoms of mental distress.

Principle components analysis. Because of these high correlations among the predictor variables, a principle components analysis was performed with a varimax rotation to further explore these relationships. The eigenvalues for the components are provided in the scree plot in Figure 1.

Figure 1. Scree plot for principle component analysis



As the scree plot indicates, one factor was prominent, accounting for 43% of the variance in the analysis; that factor was composed of the Neuroticism, OCD, GSI, and PSDI measures. Because this component is composed of all of the negative mental health measures included in the study, the component was labeled "Negative Mental Health." Components Two and Three had eigenvalues slightly above one (1.319 and 1.205, respectively) but only accounted for 16% and 15% of the variance, respectively. The other two components were not used in the analyses because their contribution to the total variance was comparatively small. In addition, both were comprised of Big 5 factors, and I was interested in examining those separately. Therefore, they were not included in this analysis.

It should be noted that the General Symptom Index (GSI) scale includes all scales from the SCL-90-R in aggregate, which means that it includes the items that make up the OCD scale. This results in the OCD items being included twice in the analysis. Although this scale is composed of relatively few scales, it was possible that this could artificially inflate the results. The principle components analysis was performed a second time without the OCD scale, and results were unchanged. Therefore, OCD was retained as part of the Mental Health scale.

Rater Bias

Generalizability theory (GT) is not commonly used in psychological research. However, it is well suited to analyses of rater bias, as was discussed in the literature review. I will provide a brief explanation of how GT is used in applications such as this, in order to provide the reader with a conceptual background for the analyses that were performed.

In typical rater bias analyses, two sources of variance, rater and target, are examined. The rater is considered a *facet* (source of error) in the terminology of GT. A standard ANOVA is used to begin the analysis; this allows the researcher to obtain sums of squares and mean squares for each source of variance in the study, as well as for the interaction between the factors (i.e., rater and target). The mean squares are then used to compute the variance estimates according to a random effects model (i.e., targets are treated as if they were sampled from a larger population of interest, and raters are treated as if they were a sample taken from a larger universe of admissible raters).

The variance component for rater ($\sigma^2(r)$) estimates the between-rater variance for all admissible raters, averaging over the population of comparable family interactions. The variance component for target ($\sigma^2(t)$) estimates the variance of observed IFIRS scores for all comparable family interactions, averaged over all admissible raters. The residual variance component ($\sigma^2(tr,e)$) is a confounded estimate of both rater-target interaction variance and error variance.

The data that were available for this project cannot be analyzed using conventional GT methods. Conventional GT techniques are only appropriate for fully crossed designs. Because this study uses an incomplete block design, alterations must be made to the standard procedures used in GT analysis. Therefore, I will use a multilevel model, which allows for the variance partitions to be obtained in a similar manner to other GT analyses. This method of data analysis has been successfully used to assess rater bias in observational data (e.g., Hoyt, 2002).

Multilevel approach

Analysis of data with a multilevel approach allows for analysis of data with complex patterns of variability, with a focus on nested sources of variability (Snijders & Bosker, 1999). Common examples of this type of data include pupils in classes or employees in companies. But this model also captures these data, which have coder nested within family and predictor variable nested within coder.

Under the umbrella of multilevel models are mixed effects models, in which it is assumed that some coefficients are fixed and others are random. Once again, this model fits these data. Therefore, these data were analyzed using a mixed effects model.

Bias Analyses

All bias analyses were performed using SAS version 8.2 for Windows, using PROC MIXED to estimate the effect the predictor variables had on each rating scale. PROC MIXED has become the standard method by which this type of variance partitioning is performed. Although other SAS procedures can be employed, PROC MIXED provides the most straightforward procedure for partitioning variance (Singer, 1998). PROC MIXED has been demonstrated to provide the same output as GENOVA (a program designed to do generalizability analyses; Snijders & Bosker, 1999) and PROC VARCOMP, both of which I used to examine rater bias in a previous study (Becker, 1999).

To verify that PROC MIXED produced equivalent results to other previously reported procedures, I analyzed a portion of these data using both PROC VARCOMP and PROC MIXED. I obtained identical results; however, PROC MIXED is significantly more straightforward and more efficient to run.

PROC MIXED provides output very similar to PROC VARCOMP. It provides an estimate of the proportion of observed variance accounted for by the different facets that are introduced into the model. First, an empty or "null" model was run for each variable of interest. Next, the analyses were repeated including the predictor variables. The output from the analysis with the predictor variables were then compared to the null model to evaluate how the proportion of explained variance changes with the inclusion of the predictor variables (e.g., personality characteristics of the coders).

The results are presented in two formats; percentage of variance accounted for, and R^2. Percentage of variance is the most commonly used method to describe results from these analyses (Singer, 1998). This is because the percentages are relatively easy to compute and easily understandable, and are useful to the consumer of research. The percentage represents the degree to which the variance attributable to coder changes with the inclusion of the predictor variable. This percentage is not provided as part of the SAS output, but is derived from a hand calculation (Singer, 1998). The formula used to compute the percentage of variance is provided in Appendix A.

The R^2 is not as commonly presented in these analyses, largely because the computation is more complicated. However, for the sake of completeness it was included for the analyses. The R^2 reports how much of the explained variance is accounted for by inclusion of the predictor variable in the model. As with the percentage of variance, SAS does not provide R^2 as part of the output. Instead, R^2 is computed from the SAS output with a hand calculation (see Snijders & Bosker, 1999). The formula used to compute R^2 is provided in Appendix A.

Results are presented in Tables 4 thru 11. Note that on scales for which there was a significant result, a value of either positive or negative is given in the "Direction" column. Positive value indicates that higher scores were associated with more rater bias; negative value indicates that lower scores are associated with more rater bias (i.e., negative value for Extraversion would indicate that lower scores on Extraversion would account for more rater bias). It should also be noted that although percentage of variance and R^2 are calculated separately, in this analysis they are typically very close and in no instance do they contradict one another.

Table 4. Hostility (HS)

Predictor	% Variance	R^2	p-Value	Direction
Neuroticism (N)	9	.0797	.3910	
Extraversion (E)	12	.1152	.0037 *	-
Openness (O)	5	.0424	.8036	
Agreeableness (A)	3	.0246	.1041	
Conscientiousness (C)	7	.0615	.0400 *	+
Obsessive-Compulsive (OCD)	5	.0416	.1114	
General Symptom (GSI)	13	.1246	.0186 *	+
Symptom Intensity (PSDI)	5	.0419	.2221	
Negative Mental Health	7	.0661	.0176 *	+

Note: * indicates significance

Table 5. Assertiveness (AR)

Predictor	% Variance	R^2	p-Value	Direction
Neuroticism (N)	0	-.0202	.6165	
Extraversion (E)	0	.0086	.8614	
Openness (O)	0	.0073	.5718	
Agreeableness (A)	0	-.0044	.7265	
Conscientiousness (C)	0	-.0306	.1908	
Obsessive-Compulsive (OCD)	0	-.0016	.9421	
General Symptom (GSI)	0	.0089	.9634	
Symptom Intensity (PSDI)	0	.0004	.9548	
Negative Mental Health	11	.1100	.5307	

Table 6. Angry Corecion (AC)

Predictor	% Variance	R^2	p-Value	Direction
Neuroticism (N)	6	.0522	.7853	
Extraversion (E)	9	.0877	.0136 *	-
Openness (O)	4	.0345	.7965	
Agreeableness (A)	3	.0293	.1241	
Conscientiousness (C)	7	.0651	.1945	
Obsessive-Compulsive (OCD)	4	.0398	.1520	
General Symptom (GSI)	11	.1007	.1236	
Symptom Intensity (PSDI)	5	.0466	.4972	
Negative Mental Health	7	.0701	.1562	

Note: * indicates significance

Table 7. Communication (CO)

Predictor	% Variance	R^2	p-Value	Direction
Neuroticism (N)	0	-.0158	.0447 *	+
Extraversion (E)	0	.0032	.0814	
Openness (O)	0	.0075	.2910	
Agreeableness (A)	0	-.0280	.1169	
Conscientiousness (C)	0	.0011	.5821	
Obsessive-Compulsive (OCD)	0	.0015	.2420	
General Symptom (GSI)	0	-.0001	.0380 *	+
Symptom Intensity (PSDI)	0	.0097	.2053	
Negative Mental Health	0	.0330	.0096 *	+

Note: * indicates significance

Table 8. Prosocial (PR)

Predictor	% Variance	R^2	p-Value	Direction
Neuroticism (N)	3	.0342	.3476	
Extraversion (E)	2	.0236	.5208	
Openness (O)	2	.0219	.5154	
Agreeableness (A)	3	.0324	.3382	
Conscientiousness (C)	3	.0345	.6245	
Obsessive-Compulsive (OCD)	2	.0198	.8436	
General Symptom (GSI)	3	.0322	.5297	
Symptom Intensity (PSDI)	1	.0119	.7947	
Negative Mental Health	7	.0742	.2785	

Table 9. Warmth/Support (WM)

Predictor	% Variance	R^2	p-Value	Direction
Neuroticism (N)	0	-.0517	.8441	
Extraversion (E)	0	-.0378	.6130	
Openness (O)	0	-.0761	.2570	
Agreeableness (A)	0	-.0366	.9719	
Conscientiousness (C)	0	-.0656	.1318	
Obsessive-Compulsive (OCD)	0	-.0430	.8798	
General Symptom (GSI)	0	-.0381	.5333	
Symptom Intensity (PSDI)	0	-.0366	.9391	
Negative Mental Health	0	-.2241	.7696	

Table 10. Antisocial (AN)

Predictor	% Variance	R^2	p-Value	Direction
Neuroticism (N)	4	.0379	.0007 *	+
Extraversion (E)	12	.1168	.0002 *	-
Openness (O)	0	-.0630	.6313	
Agreeableness (A)	0	-.0259	.4754	
Conscientiousness (C)	2	.0161	.3150	
Obsessive-Compulsive (OCD)	0	-.0154	.0289 *	+
General Symptom (GSI)	10	.0953	.0001 *	+
Symptom Intensity (PSDI)	0	-.0242	.0808	
Negative Mental Health	32	.3084	.0001 *	+

Note: * indicates significance

Table 11. Listener Responsiveness (LR)

Predictor	% Variance	R^2	p-Value	Direction
Neuroticism (N)	1	.0133	.0013	
Extraversion (E)	5	.0448	.0012 *	-
Openness (O)	0	-.0215	.2312	
Agreeableness (A)	0	-.0152	.0797	
Conscientiousness (C)	4	.0405	.1743	
Obsessive-Compulsive (OCD)	2	.0187	.0632	
General Symptom (GSI)	3	.0256	.0007 *	+
Symptom Intensity (PSDI)	0	.0019	.1071	
Negative Mental Health	6	.0606	.0003 *	+

Note: * indicates significance

Big 5 results

Extraversion was significantly related to rater bias across several rating scales. It accounted for 12% of variance attributable to coder on the Hostility scale; 9% on the Angry Coercion scale; 12% on the Antisocial scale; and 5% on the Listener Responsiveness scale. In each case, the direction of the relationship was negative. That is, as extraversion goes down, the amount of bias introduced into the data goes up. In other words, coders who are introverted introduce more bias than those who are extraverted.

The Neuroticism scale accounted for 4% of bias on the Antisocial scale, and 1% on the Listener Responsiveness scale. Both of these are in the positive direction – as neuroticism goes up, so does rater bias. There were no significant results for the scales of Openness, Agreeableness, and Conscientiousness.

Mental health results

The GSI scale (overall mental health of coder) accounted for 13% of variance attributable to coder on the Hostility scale, 10% on the Antisocial scale, and 3% on the Listener Responsiveness scale. All were in the positive direction, indicating that greater mental distress predicted greater rater bias.

The PSDI scale (how intense coders' psychological symptoms are) was not significantly related to rater bias on any rating scales. This indicates that the intensity of coders' symptoms did not account for rater bias.

The composite Mental Health scale accounted for 7% of the variance attributable to coder on the Hostility scale, 32% on the Antisocial scale, and 6% on the Listener Responsiveness scale. Once again, all were in the positive direction, indicating greater composite mental distress predicted greater rater bias.

Summary of Results

Two trends were evident from these results. First, coders who were introverted introduced more rater bias than those who were extroverted. Second, coders who experienced poor mental health introduced more rater bias than coders who were psychologically healthy. These findings were generally true with respect to scales measuring negative or neutral attributes of the targets. The effects were not found on scales measuring positive attributes of the target.

DISCUSSION

In this discussion, I will first examine deficits in the social science literature that this study attempted to address. Second, I will discuss the implications of the results of this research. Next, I will provide suggestions for the selection and training of observational coders in order to minimize rater bias. Finally, I will end with directions for future research.

Methodological Deficits

Because so few studies have been conducted that examine predictors of rater bias, any methodological limitations of those studies that make the results more difficult to use are damaging. For example, some previous studies examining predictors of rater bias have indicated a specific personality variable that seems to explain some of the rater bias that is observed. However, these studies often do not discuss the amount of bias these facets account for, or sometimes even the direction of the effect (e.g., is higher agreeableness associated with more or less bias?). The lack of reports concerning magnitude or direction of effects for predictor variables makes it very difficult to know if the variable of interest is truly one that provides explanatory power regarding rater bias. In this research I used statistical techniques that allowed an examination of both the magnitude and direction of the effect of the predictor variables. These results (i.e., magnitude and direction) are of much greater utility to researchers attempting to understand what facets account for rater bias.

This area has also suffered from researchers' use of questionable methods to assess personality or mental health facets of the rater. These questionable practices include questionnaires generated by researchers that have not been validated against established instruments. I chose to use two standardized, established assessment tools (i.e., the SCL-90-

R and NEO-PI-R) that have been demonstrated to be reliable and valid. The use of these instruments has the advantage of giving us greater confidence that we have accurately measured the variables of interest, and gives us a commonly understood definition of what we are measuring. This eliminates the "apples and oranges" phenomenon that occurs when each researcher has his or her own definition of a variable of interest.

Many studies examining rater bias use coding systems developed for a specific rater bias project, and often these coding systems are flawed. This diminishes our ability to understand if the observed bias is truly attributable to the coder or due to some flaw of the coding system. By using coding data that were collected using a well-established coding system with demonstrated utility, I was better able to establish that observed bias was attributable to the coder and not to a flawed coding system.

Finally, it has been noted that the majority of studies in this area use very small samples of coding data from which to draw bias estimates. This is understandable, given the large amount of resources required to obtain coding data. However, by using small samples of coding data it is much harder to establish the existence of rater bias with any degree of confidence, given that the data can be skewed by one outlier target. This research project drew from a vast longitudinal database, allowing the inclusion of over 100,000 individual data points. This allowed me to establish with confidence the degree of rater bias that was present.

Implications of the Results

When looking at the results of this research, it is helpful to first remember that the coding data taken from the IFIRS represents a "best case" scenario. The observational

coders working with this system have received an extensive amount of training, and receive ongoing training on a weekly basis. They meet with other raters that have coded the same tasks in order to decide on "correct" scores, which provides coders the opportunity for recalibration. The IFIRS coding system has been operationalized and refined over several years, and has demonstrated utility. Despite training procedures that far exceed those used by most researchers to ensure accurate coding data, rater bias does indeed exist in these data. If any system could "train out" rater bias, it would be the IFIRS. The amount of bias observed in these data could be much higher in settings that provide less thorough training, ongoing supervision, and re-calibration.

Two general trends were evident in the findings from these analyses. First, coders who were more introverted tended to introduce more rater bias into the data. Second, coders who were suffering from some type of mental health issue introduce more rater bias than those in better mental health.

Introversion

This research has demonstrated that coders who are introverted introduce more rater bias than those who are extraverted. This is consistent with the Ambady et al. (1995) finding that coders that rated themselves as quieter and less socially skilled were found to be "worse" coders.

This result is intuitively logical. Coding of family interactions involves observing social interactions between multiple targets, and applying meaning to the behaviors that are displayed. Possessing social and communication skills, as well as being able to use those skills comfortably and confidently, helps provide insight and understanding of the tasks that

are observed. This insight into social interactions can then be used to understand the dynamics at play between targets.

This insight would likely not be as present for those coders with poor social skills, or those who use their skills sporadically. The general internal orientation and discomfort with social interaction associated with introversion seems to affect the coders' ability to accurately capture the behaviors of others.

The fact that introversion provides explanatory power on those codes that reflect negative (but not positive) attributes of the targets is interesting. As noted earlier, the Extraversion scale assesses sociability, gregariousness, and assertiveness. The target behaviors measured by the negative attribute scales of the IFIRS (including hostile and antisocial actions toward others) would likely feel very intense and threatening to someone who is introverted, and therefore they may inflate those scores beyond what is accurate. When viewed from this perspective, it is logical that those who are less sociable, gregarious and assertive would find the behaviors captured in scales that measure negative facets of the targets foreign or distasteful, and therefore more rater bias could be introduced. However, behaviors reflecting positive attributes of the coders would be more common and less threatening, and therefore the coders' introversion would not have the same level of impact.

Mental health

The results of this study indicate that coders who have emotional or mental health difficulties introduce more rater bias than those who report good mental health. There are several possible explanations for this effect. First, coders suffering from poor mental health may have greater difficulty applying the coding system consistently or correctly. Those with

poor mental health demonstrate impaired functioning in multiple areas of their lives, including work (Reid, Ballis, & Sutton, 1997). Decreased concentration and distractibility are common. Also, motivation can decrease, and therefore coders might not be as likely to try to code accurately as those who are in better mental health.

However, this explanation (i.e., that coders are not functioning as well, and therefore not coding as accurately) would predict that bias would be introduced across all scales, and that was not the case. It is commonly reported that persons suffering from mental illnesses such as depression view the world from a negative perspective; they see the negative around them that they are feeling internally (McNamara, 1992). When viewed from this perspective, it follows that coders who are suffering from mental health challenges may focus too heavily on the negative behaviors exhibited in the tasks being observed, while at the same time minimizing the positive behaviors that are exhibited. This would explain the inflation introduced into the negative scales but not into the positive scales.

The explanation that coders who have a negative state of mind at the time of coding can cause them to focus more heavily on negative behaviors of targets is not new. Indeed, many researchers have demonstrated the phenomenon that coders in a "bad" mood are harsher in their ratings (e.g., Hampson, 1984). While I was not examining state-dependent mood but rather a more global state of mental health, the results are consistent with the view that coders in a negative state will not code as reliably as would be desired, and will artificially inflate scores on negative codes.

Finally, it should be noted that the fact that no personality or mental health factors were predictive of rater bias on scales measuring positive attributes of the target could indicate that those scales are simply not as prone to bias from personality or mental health

variables. Although rater bias is present in these scales, it could be coming from other unmeasured sources.

Selection of Coders

This research demonstrated that as introversion and poor mental health increased, rater bias on scales measuring negative attributes of the targets also increased. This finding suggests that by controlling for the impact of introversion and mental health, researchers might lower the amount of bias present in their data, and improve their ability to generalize from those data.

The easiest way to achieve this control would be to only employ coders that fit the "ideal" template (i.e., high extraversion, low mental health difficulties). Researchers could use prescreening measures when hiring coders to select only those with the desirable traits of extraversion and good mental health (i.e., low neuroticism). This prescreening could be done with standardized assessment tools (such as those used in this research), or assessment tools could be developed to assess potential coders on the facets of extraversion and mental health only.

However, prescreening during hiring might not be possible; for example, researchers might not have a large enough pool of candidates to be able to be this selective. If prescreening of new coders is not possible, researchers could adjust training procedures to help those who are introverted or who score higher on negative mental health code negative behaviors of targets accurately. This might involve spending more time on training for these scales, or requiring special testing only on these scales to ensure that the coder is able to accurately capture the behaviors being observed.

It is possible that there could be multiple "ideal" coders. Extroversion and mental health only account for a portion of the rater bias present in these data. The majority of rater bias that was observed remains unexplained. There are a multitude of individual difference factors that could further explain the origin of rater bias, including other intrapsychic factors (e.g., intelligence), demographic factors (e.g., race), and state-dependent factors (e.g., mood). Until other individual differences can be examined to see what impact they have on rater bias, we will not be able to identity the ideal coder with confidence.

Directions for Future Research

This is the first study of this type to be conducted. Replication is necessary to ensure that the findings are generalizable. In particular, replication with a larger sample of coders would increase statistical power.

The coder sample in this research was very homogeneous; all coders were Caucasian and college-educated. The majority were women. Most coders were from Iowa and were married with children. Whether these results would be present in groups of coders representing different demographic or racial groups is unknown. Research cited in the literature review demonstrated gender differences in rater bias studies; this would indicate the need to replicate these findings with a larger group of male coders.

The IFIRS coding system is not typical of the majority of observational coding systems currently in use. IFIRS is designed to minimize the introduction of variance not attributable to the target as much as possible; its 351-page manual provides lengthy definitions of codes, and extensive examples of the types of behaviors being coded. The IFIRS system has been through four previous editions, each of which built on the last to

better operationalize the behavioral characteristics of the targets. I could not identify any other coding systems currently in use that are as comprehensive as the IFIRS. If the majority of coding systems being used are not as specific and extensive as the IFIRS, these results might not be a good analog to what would be found in other coding systems. Therefore, further research in this area should use a more "typical" coding system to examine if these results are replicable.

Coders in this study received approximately 180 hours of training before ever coding observational tasks for the ISBR data set. In additionally, they receive ongoing training throughout their employment at ISBR. This is significantly more training than any other coding system I encountered in the literature. Research has noted that increased training tends to reduce rater bias (Hoyt, 2000). Given this finding, it would be expected that higher levels of rater bias would be present in data gathered by coders who have not been as extensively trained. Replicating this research with a pool of coders who received less training could yield results that are more applicable to other researchers using observational coding systems.

Very little research exists that attempts to explain the origins of rater bias. Personality and mental health were examined in this study because previous research has indicated that these factors might impact rater bias. This was demonstrated to be true. However, only a portion of the observed rater bias was explained by these facets. These results indicate that there may be other factors that account for rater bias in observational data. Future researchers should continue to explore a variety of facets attributable to the coder, or to the coder-target pair, that could help explain the causes of rater bias.

Summary

It is often stated in the literature that observational coding is the best way to assess the behavior of others. We are taught that by using observational data we can avoid a plethora of problems that are inherent in other types of data collection methods. For example, one text in the social sciences points out flaws in multiple types of data collection methodologies (e.g, self-report) but after stating that observational coding is "highly valid and accurate" the authors indicate that the only deficit of observational coding is that the data are "laborious to collect" (Zebrowitz, 1990, p. 76). No mention is made of the issue of rater bias, or the impact that individual differences among the coders can have on the accuracy of their coding work. Given the degree to which rater bias can attenuate results, such bias is to be avoided at all costs. The fact that the issue of rater bias is so often ignored by researchers calls into question the results of many published studies.

By demonstrating the degree to which introversion and poor mental health explain rater bias, I have attempted to provide researchers using observational coding with information they can use to help control the impact of these coder characteristics. Being able to control the amount of rater bias in observational data will lead to more accurate data, and in turn increases our confidence in results generated from these data.

Given these findings, it could be appropriate for those using observational coding to prescreen potential coders, selecting those demonstrating extroversion and good mental health. Alternatively, training methods could be developed to specifically address the impact of these facets, in an effort to reduce their impact on observational data.

Ultimately, further research will be required before we can identity the "ideal coder." It is my hope that this research will be pursued by others in the future, because the potential

benefits of explaining and controlling rater bias will have a positive impact on countless future studies in the social sciences.

APPENDIX A: COMPUTATIONAL FORMULAS

Below are the formulas used for the hand calculations noted in the Results section.

Percentage of variance:

$$\% = \frac{\tau^2_{m1} - \tau^2_{m0}}{\tau^2_{m1}}$$

See Singer, 1998

R-square:

$$R^2 = \left(1 - \frac{\sigma^2_{m1} + \tau^2_{m1}}{\sigma^2_{m0} + \tau^2_{m0}}\right)$$

See Snijders & Bosker, 1999

APPENDIX B: INFORMED CONSENT FORM

INFORMED CONSENT DOCUMENT

Title of Study: **Individual Differences and Perception**
Investigator: **Mark R. Becker, M.S.**
2223 Student Services Building, 3rd Floor
(515) 294-0156
mrbecker@iastate.edu

This is a research study. Please take your time in deciding if you would like to participate. Please feel free to ask questions at any time.

INTRODUCTION

The purpose of this study is to learn about how different personality variables affect the way people observe others. You are being invited to participate in this study because you are currently a Family Interaction Analyst.

DESCRIPTION OF PROCEDURES

If you agree to participate in this study, your participation will last for approximately one hour. During the study you may expect the following study procedures to be followed: you will be asked to complete two questionnaires that ask how closely you feel a series of statements describes you. You may skip any question that you do not wish to answer or that makes you feel uncomfortable.

RISKS

While participating in this study you may experience the following risks: because you are answering questions about yourself and your personality, you could experience some discomfort if questions cause you to reflect on facets of yourself that are unpleasant to you. There are no anticipated physical risks.

BENEFITS

If you decide to participate in this study there may be no direct benefit to you. It is hoped that the information gained in this study will benefit society by helping us better understand how people perceive each other.

ALTERNATIVES TO PARTICIPATION

If you chose not to participate, you may substitute coding hours for the time spent in the general meeting completing the questionnaire.

COSTS AND COMPENSATION

You will not have any costs from participating in this study.

PARTICIPANT RIGHTS

Your participation in this study is completely voluntary and you may refuse to participate or leave the study at any time. If you decide to not participate in the study or leave the study early, it will not result in any penalty or loss of benefits to which you are otherwise entitled.

CONFIDENTIALITY

Records identifying participants will be kept confidential to the extent permitted by applicable laws and regulations and will not be made publicly available. However, federal government regulatory agencies and the Institutional Review Board (a committee that reviews and approves human subject research studies) may inspect and/or copy your records for quality assurance and data analysis. These records may contain private information.

To ensure confidentiality to the extent permitted by law, the following measures will be taken: no names will be attached to any data, and coder ID number will NOT be used to identify you; you will be identified only by a randomly selected number. Although signed informed consent is being obtained, this form and your questionnaires will be collected and stored separately, and the two cannot be paired. The primary investigator (Mark Becker) will

be the only person with access to the raw data. No employee of the Institute for Social and Behavioral Research will have access to these data, or to any other data which might identify you. If the results are published, your identity will remain confidential.

Additionally, it is important that you understand that this is research being conducted through the Iowa State University Department of Psychology, and NOT through the Institute for Social and Behavioral Research. Therefore, no Institute personnel will have access to your personal information now or in the future, and all data will be analyzed at the group level.

QUESTIONS OR PROBLEMS

You are encouraged to ask questions at any time during this study. For further information about the study contact Mark Becker, 2223 Student Services Building, (515) 294-0156; mrbecker@iastate.edu . You may also contact the project supervisor, Carolyn Cutrona, W159 Lagomarcino Hall, (515) 294-6784; ccutrona@iastate.edu. If you have any questions about the rights of research subjects or research-related injury, please contact the Human Subjects Research Office, 2810 Beardshear Hall, (515) 294-4566; meldrem@iastate.edu or the Research Compliance Officer, Office of Research Compliance, 2810 Beardshear Hall, (515) 294-3115; dament@iastate.edu.

****

PARTICIPANT SIGNATURE

Your signature indicates that you voluntarily agree to participate in this study, that the study has been explained to you, that you have been given the time to read the document and that your questions have been satisfactorily answered. You will receive a copy of the signed and dated written informed consent prior to your participation in the study.

Participant's Name (printed) _____

_____ _____

(Participant's Signature) (Date)

INVESTIGATOR STATEMENT

I certify that the participant has been given adequate time to read and learn about the study and all of their questions have been answered. It is my opinion that the participant understands the purpose, risks, benefits and the procedures that will be followed in this study and has voluntarily agreed to participate.

_____ _____

(Signature of Person Obtaining Informed Consent) (Date)

APPENDIX C: DEBRIEFING FORM

Individual Differences and Perception

Thank you for participating in today's research study!

In this study, we are trying to understand how different individual differences impact the way in which people perceive others. By taking part, you have provided valuable assistance in helping us understand this area more completely.

It is unlikely that you will experience any discomfort or distress from filling out these questionnaires. If, however, these questionnaires raise personal issues for you that you would like to discuss further, you are encouraged to call The Richmond Center at (515) 232-5811.

If you have any questions or concerns regarding this research, you may speak with Mark Becker. He will be available after today's testing to discuss this research with you further. He can also be reached at (515) 294-0156, or at mrbecker@iastate.edu.

ACKNOWLEDGEMENTS

I wish to thank Todd Abraham, Iowa State University Department of Psychology, for his consultation on the SAS programming code used in these analyses.

REFERENCES

Aiken, L. R. (1999). *Personality assessment: Methods & practices* (3rd ed., revised). Kirkland, WA: Hogrefe & Huber Publishers.

Alexander, J. F., Newell, R. M., Robbins, M. S., & Turner, C. W. (1995). Observational coding in family therapy process research. *Journal of Family Psychology, 9*, 355-365.

Ambady, N., Hallahan, M. & Rosenthal, R. (1995). On judging and being judged accurately in zero-acquaintance situations. *Journal of Personality and Social Psychology, 69,* 518-529.

Baskett, G. D. (1973). Interview decisions as determined by competency and attitude similarity. *Journal of Applied Psychology, 57*, 343-345.

Basow, S. A., & Rubenfeld, K. (2003). "Troubles Talk"; Effects of gender and gender-typing. *Sex Roles, 48*, 183-187.

Becker, M. R. (1999). *Rater bias in observational data: A generalizability analysis.* Unpublished master's thesis, Iowa State University, Ames.

Bernieri, F. J., & Gillis, J. S. (1995). Personality correlates of accuracy in a social perception task. *Perceptual and Motor Skills, 81*, 168-170.

Bettencourt, B. A., Dill, K. E., Greathouse, S. A., Charlton, K., & Mullholland, A. (1997). Evaluations of ingroup and outgroup members: The role of category-based expectancy violation. *Journal of Experimental Social Psychology, 33*, 244-275.

Borman, W. C. (1975). Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. *Journal of Applied Psychology, 60*, 556-560.

Bower, G. H. (1981). Mood and memory. *American Psychologist, 36*, 129-148.

Brennen, R. L. (1992). *Elements of generalizability theory.* Iowa City, IA: ACT Publications.

Coates, T. J., & Thoresen, C. E. (1978). Using generalizability theory in behavioral observation. *Behavior Therapy, 9,* 605-613.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Colvin, C. R., & Bundick, M. J. (2001). In search of the good judge of personality: Some methodological and theoretical concerns. In J. Hall & F. Bernieri (Eds.), *Interpersonal sensitivity: Theory and measurement* (pp. 47-65). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Cook, E. P. (1985). *Psychological androgyny.* New York: Pergamon Press.

Cook, M. (1979). *Perceiving others.* New York: Methuen & Company.

Cook, M. (1989). *Perceiving others: The psychology of interpersonal perception.* New York: Methuen & Company.

Costa, P. T. & McCrae, R. R. (1986). Personality stability and its implications for clinical psychology. *Clinical Psychology Review, 6,* 407-423.

Costa, P. T., & McCrae, R. R. (1992). *NEO PI-R professional manual.* Odessa, FL: Psychological Assessment Resources.

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

Davis, H. M., & Kraus, A. L. (1997). Personality and empathic accuracy. In W. Ickes (Ed.), *Empathic accuracy* (pp. 145-165). New York: Guilford Press.

Davison, G. C., & Neale, J. M. (1994). *Abnormal psychology* (6th ed.). Oxford, England: John Wiley & Sons.

De Raad, B. (2000). *The big five personality factors: The psycholexical approach to personality*. Kirkland, WA: Hogrefe & Huber Publishers.

Derogatis, L. R. (1997). *SCL-90-R Administration, Scoring, and Procedures Manual* (3rd ed.). Minneapolis: NCS Pearson, Inc.

Elliot, R. (1991). Five dimensions of therapy process. *Psychotherapy Research, 1*, 92-103.

Elmes, D. G., Kantowitz, B. H., & Roediger, H. L. (1992). *Research Methods in Psychology*. (4th ed.). St. Paul, MN: West Publishing Company.

Feeley, T. H. (2002). Evidence of halo effects in student evaluations of communication instruction. *Communication Education, 51*, 225-236.

Feinberg, M., Neiderhiser, J., Howe, G., & Hetherington, E. M. (2001). Adolescent, parent, and observer perceptions of parenting; Genetic and environmental influences on shared and distinct perceptions. *Child Development, 72*, 1266-1284.

Floyd, F. J., Markman, H. J. (1983). Observational biases in spouse observation: Toward a cognitive/behavioral model of marriage. *Journal of Consulting and Clinical Psychology, 51*, 450-457.

Floyd, F. J. (1989). Segmenting interactions: Coding units for assessing marital and family behaviors. *Behavioral Assessment, 11*, 23-29.

Frank, K., & Anderson, O. (2004). *Summary of published works that include observational measures from the Iowa Family Interaction Rating Scales* (Tech. Rep. No. 2). The University of Tennessee at Knoxville, Family Life Project.

Funder, D. C., & Colvin, C. R. (1997). Congruence of others' and self-judgments of personality. In R. Hogan, J. Johnson, & S. Briggs (Eds.), *Handbook of personality* (pp. 617-647). San Diego, CA: Academic Press.

Funder, D. C. (1999). *Personality judgment: A realistic approach to person perception.* San Diego, CA: Academic Press.

Greenburg, L. S. (1995). The use of observational coding in family therapy research: Comment on Alexander et al. (1995). *Journal of Family Psychology, 9*, 366-370.

Gottman, J. M. (1979). *Marital interaction.* New York: Academic Press.

Groth-Marnat, G. (1997). *Handbook of psychological assessment* (3rd ed.). New York: John Wiley & Sons, Inc.

Hampson, S. E. (1984). Personality traits: in the eye of the beholder or the personality of the perceived? In M. Cook (Ed.)., *Issues in perception* (pp. 28-47). New York: Methuen, Inc.

Hersen, M., & Turner, S. M. (1994). *Diagnostic interviewing* (2nd ed.). New York: Plenum.

Heyman, R. E. (2001). Observation of couple conflicts: Clinical assessment applications, stubborn truths, and shaky foundations. *Psychological Assessment, 13*, 5-35.

Hill, C. E., Helmes, J. E., Tichenor, V., Spiegel, S., O'Grady, K. E., & Perry, E. (1988). Effects of therapist response modes in brief psychotherapy. *Journal of Counseling Psychology, 35*, 222-233.

Hill, C. E., O'Grady, K. E., & Price, P. (1988). A method for investigating sources of rater bias. *Journal of Counseling Psychology, 35*, 346-350.

Hill, C. E., & Stephany, A. (1990). The relationship of nonverbal behaviors to client reactions. *Journal of Counseling Psychology, 37*, 22-26.

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods, 5*, 64-86.

Hoyt, W. T. (2002). Bias in participant ratings of psychotherapy process: An initial generalizability study. *Journal of Counseling Psychology, 49*, 35-46.

Hoyt, W. T., & Kerns, M. D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods, 4*, 403-424.

Hoyt, W. T., & Melby, J. N. (1999). Dependability of measurement in counseling psychology: An introduction to generalizability theory. *Counseling Psychologist, 27*, 325-352.

John, O. P., & Robins, R. W. (1994). Accuracy and bias in self-perception: Individual differences in self-enhancements and narcissism. *Journal of Personality and Social Psychology, 66*, 206-219.

Johnson, J. A. (2000). Predicting observers' ratings of the Big Five from the CPI, HPI, and NEO-PI-R: A comparative validity study. *European Journal of Personality, 14*, 1-19.

Kahneman, D. & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review, 93*, 136-153.

Kenny, D. A., Albright, L., Malloy, T. E., & Kashy, D. A. (1994). Consensus in interpersonal perception: Acquaintance and the Big Five. *Psychological Bulletin, 116*, 245-258.

King, K. (2001). A critique of behavioral observational coding systems of couples' interaction: CISS and RCISS. *Journal of Social and Clinical Psychology, 20*, 1-23.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19*, 3-31.

Lippa, R. A., & Dietz, J. L. (2000). The relation of gender, personality and intelligence to judges' accuracy in judging strangers' personality from brief video segments. *Journal of Nonverbal Behavior, 24*, 25-43.

Mahalik, J. R.., Hill, C. E., O'Grady, K. E., & Thompson, B. J. (1993). Rater characteristics influencing rating on the Checklist of Psychotherapy Transactions-Revised. *Psychotherapy Research, 3*, 47-56.

Margolin, G., Hattem, D., John, R., & Yost, K. (1985). Perceptual agreement between spouses and outside observers when coding themselves and a stranger dyad. *Behavioral Assessment, 7*, 235-247.

Markman, H. J., Leber, D., Cordova, A. D., & St. Peters, M. (1995). Behavioral observation and family psychology -- strange bedfellows or happy marriage?: Comment on Alexander et al. (1995). *Journal of Family Psychology, 9*, 371-379.

Markman, H. J., & Notarius, C. (1987). Coding marital and family interaction. In T. Jacob (Ed.)., *Family Interaction and Psychopathology*. New York: Plenum.

McNamara, K. (1992). Depression assessment and intervention: current status and future directions. In S. Brown & R. Lent (Eds.)., *Handbook of Counseling Psychology* (2nd ed.). New York: John Wiley & Sons, Inc.

Melby, J., Conger, R., Book, R., Rueter, M., Lucy, L., Repinski, D., Rogers, S., Rogers, B., Scaramella, L. (1998). *The Iowa Family Interaction Rating Scales* (5th ed.). Unpublished manuscript. Institute for Social and Behavioral Research, Iowa State University, Ames.

Melby, J. N., & Lorenz, F. O. (1996, June). *Conceptual and methodological considerations in the use of observational data.* Poster session presented at the Family Research Consortium Summer Institute, San Diego.

Moskowitz, D. S. (1986). Comparison of self-reports, reports by knowledgeable informants, and behavioral observation data. *Journal of Personality, 54*, 294-317.

Petkova, E., Quitkin, R., McGrath, P, Stewart, J., & Klein, D. (2000). A method to quantify rater bias in antidepressant trials. *Neuropsychopharmacology, 22*, 559-565.

Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin, 28*, 369-381.

Reid, W. H., Balis, G. U., & Sutton, B. J. (1997). *The treatment of psychiatric disorders* (3rd ed.). Britol, PA: Brunner/Mazel Publishers.

Robinson, E. A., & Price, R. G. (1980). Pleasurable behavior in marital interaction: an observational study. *Journal of Consulting and Clinical Psychology, 48*, 117-118.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*, 413-428.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 24*, 323-355.

Smith, E. R. (1991). The role of exemplars in social judgment. In L. L. Martin & A. Tesser (Eds.), *The construction of social judgments* (pp. 7 – 102). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* London: SAGE Publications, Ltd.

Swim, J. K. (1994). Perceived versus meta-analytic effect sizes: An assessment of the accuracy of gender stereotypes. *Journal of Personality and Social Perception, 66*, 21-36.

van der Valk, J. C., van den Oord, E. J. C. G., Verhulst, F. C., & Boomsma, I. (2001). Using parental ratings to study the etiology of 3-year-old twins' problem behaviors: Different views or rater bias? *Journal of Child Psychology & Psychiatry, 42*, 921-931.

Winquist, L. A., Mohr, C. D., & Kenny, D. A. (1998). The female positivity effect in the perception of others. *Journal of Research in Personality, 32*, 370-388.

Whitley, B. E. (1996). *Principles of research in behavioral science.* Mountain View, CA: Mayfield Publishing Company.

Wyer, R. S., Lambert, A. J., Budesheim, T. L., & Gruenfeld, D. H. (1991). Theory and research on person impression formation: A look to the future. In L. L. Martin & A. Tesser (Eds.), *The construction of social judgments* (pp. 7 – 102). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Zebrowitz, L. A. (1990). *Social perception.* Buckingham, England: Open University Press.