Improving Assessment of Personality Disorder Traits Through Social Network Analysis

Allan Clifton[1], **Eric Turkheimer**[2], and **Thomas F. Oltmanns**[3]

[1]Vassar College

[2]University of Virginia

[3]Washington University, St. Louis

Abstract

When assessing personality disorder traits, not all judges make equally valid judgments of all targets. The present study uses social network analysis to investigate factors associated with reliability and validity in peer assessment. Participants were groups of military recruits ($N = 809$) who acted as both targets and judges in a round-robin design. Participants completed self- and informant versions of the Multisource Assessment of Personality Pathology. Social network matrices were constructed based on reported acquaintance, and cohesive subgroups were identified. Judges who shared a mutual subgroup were more reliable and had higher self-peer agreement than those who did not. Partitioning networks into two subgroups achieved more consistent improvements than multiple subgroups. We discuss implications for multiple informant assessments.

In both research and clinical settings, personality disorders (PDs) are most often diagnosed on the basis of self-report, obtained through written inventories or clinical interview. However, information gathered from self-report tends to differ substantially from how others view the individual (Clifton, Turkheimer, & Oltmanns, 2004). A review of 30 published studies of self- and informant reports of personality disorders concluded that self-informant correspondence was "modest at best" (Klonsky, Oltmanns, & Turkheimer, 2002, p. 308) with a median correlation of $r = .36$ in studies of DSM personality disorders.

Peer perceptions of pathological personality traits are usually obtained from a knowledgeable informant who describes the personality of the participant via questionnaire or structured interview (e.g., McCrae & Costa, 1987). However, informants selected by the participant may suffer from what has been described as the "letter of recommendation" problem (Klonsky et al., 2002). That is, a close friend, spouse, or relative chosen as an informant may describe the participant in an overly positive light. Unselected peers who interact with the individual on a regular basis, such as coworkers or classmates, may be less biased in their judgments.

Correspondence concerning this article should be addressed to Allan Clifton, Department of Psychology, Vassar College, Box 127, 124 Raymond Ave., Poughkeepsie, NY 12604-0127. alclifton@vassar.edu.

To date, relatively few studies of personality pathology have incorporated information from multiple unselected peers. Fiedler, Oltmanns, and Turkheimer (2004) administered self- and peer-report measures of PDs to 1,080 military recruits and followed them prospectively for 2 years. The authors found that both self- and peer report provided incremental validity in predicting maladaptive functioning (e.g., early discharge from the military). Similarly, Clifton, Turkheimer, and Oltmanns (2005) examined the relationship between self- and peer-reported personality disorder traits and interpersonal problems in 393 college undergraduates. Canonical analyses found that self- and peer sources described a similar relationship between pathological traits and interpersonal behavior but identified completely different individuals as manifesting such behaviors. These findings emphasize the importance of obtaining information from multiple sources, rather than relying solely on self-report.

A possible drawback to the use of multiple unselected peers is the complicating effect of group dynamics. Ratings in large group studies may be affected by a variety of interpersonal variables, including rater-rater acquaintance, rater-target acquaintance, degree of overlap in observations by raters, ingroup and outgroup effects, differing interpretations of behavior by judges, differing average rating tendencies by judges, and a host of other factors (e.g., Tajfel, 1978; Kenny, 1994; Kenny & Kashy, 1994; Park & Judd, 1989).

In laboratory studies, or studies of individual informants, factors such as these may be more easily taken into account, as informants can be interviewed in depth regarding their association with the target. For example, given perceivers previously unacquainted with the target, behavioral perception overlap may be manipulated by allowing perceivers to observe particular subsets of the target's behavior (e.g., Kenny, Albright, Malloy, & Kashy, 1994). Similarly, given perceivers previously unacquainted with one another, communication among judges could be manipulated by circumscribing the amount or type of communications they engage in. These factors become more difficult to interpret in studies of larger, preexisting groups. When participants have previously known one another for significant amounts of time, they may have interacted in any number of settings and situations, creating an unknown amount of similarity among perceivers'; interactions with the target. The amount of communication among two perceivers regarding any given target is similarly difficult to ascertain. This is not to say that such information is impossible to obtain but simply that obtaining it requires more effort than most researchers can afford. To evaluate the effects of communication in large groups of acquaintances requires that each judge be asked about his or her communication with each other judge regarding each target. In a round-robin design of 40 participants, this would involve $(40 \times 39 \times 38) = 59{,}280$ additional pieces of information, all of which are retrospective estimates by participants, with the additional error entailed (e.g., Bernard, Killworth, Kronenfeld, & Sailer, 1984). Assessing perception overlap and shared meaning systems in perceivers is equally difficult and time consuming.

Despite these limitations, treating all judges in a group as equally good informants, or accounting only for judge-target acquaintance, may overlook valuable sources of information (Kenny, 1994). Traditional assessments of personality perception have generally focused on dyads, in which one or two informants rate the personality of a target person. In

contrast, social network analysis (SNA) looks at all of the targets and judges within an entire social system, focusing on the patterns of relationships (Kanfer & Tanaka, 1993). Information obtained from social network analysis of the data provides a simplified measure of a highly complex interpersonal environment. Kanfer and Tanaka (1993) noted that social network analysis can "address fundamental questions about the social nature of personality constructs, including the perception of self and others" (pp. 735–736). Although social network analysis has been widely used in a variety of related disciplines, it has rarely been applied to personality assessment (Funder & West, 1993) and never to personality pathology.

SNA analyzes the patterns of relationships within an entire bounded social network. Such a network might be defined as the employees in a workplace, the members of a village, students in a sorority, or any other group of individuals around which a meaningful boundary can be drawn. Given this bounded network, SNA can be used to interpret an individual's relationship to the entire group, rather than just his or her relationship to another individual. For example, within a network, cohesive subgroups of individuals with strong interrelations to one another can often be identified. Previous findings suggest that an individual's position in a social network is likely to be associated with both the way that individual perceives other members of the network (e.g., Breiger & Ennis, 1979) and the way in which that individual's personality is perceived by others (e.g., Michaelson & Contractor, 1992). Therefore, in order to understand the relationship between self- and peer perceptions of personality disorders, a better understanding of the impact of the social network is essential. Taking social network structures into account could provide a parsimonious way of increasing both reliability and validity in peer evaluations, thereby improving our understanding of the relationship between self- and peer perceptions of personality pathology.

The present study investigates the relationship between network structure and self- and peer reports of personality pathology in 21 large groups of peers. Specifically, two questions of interest were investigated: First, how does judge consensus within subgroups compare with judge reliability between subgroups? We expect that reports made by peers within a common subgroup will be more similar than those made between subgroups. And second, how does self-peer agreement within subgroups compare to that between subgroups? We hypothesize that peers' reports of those within their cohesive subgroups will demonstrate greater agreement with self-report than do evaluations made between subgroups.

Methods

Participants

Participants (N = 809, 533 male, 276 female) were Air Force recruits who were assessed at the end of 6 weeks of basic training at Lackland Air Force Base. The present sample is a subset of a larger sample, described more fully in Oltmanns and Turkheimer (2006). The current study does not overlap with the sample described in Fiedler et al. (2004) but comprises the most recently assessed participants, for whom a measure of acquaintance was added, as described below.

The participants in our sample were enlisted personnel who would eventually receive assignments as military police, mechanics, computer technicians, or other supportive roles. Their mean age was 20 years (*SD* = 5), and 99% were high school graduates. Sixty-four percent described themselves as white, 16% as black, 4% as Asian, 4% as biracial, 1% as Native American, and 12% as another racial group. Air Force recruits undergo mandatory psychological screenings before beginning basic training in order to screen out those with Axis I psychopathology. These screenings, however, were not designed to detect or screen out those with Axis II personality disorders.

The participants were members of 21 "flights," groups of 27 to 54 recruits who went through training together. Six of these flights were single-gender male flights, and 15 were mixed-gender flights (see Table 1). Recruits in a given flight spend nearly 24 hours a day together, including time training, eating, and sleeping. Recruits'; names are written on their uniforms and are used frequently by their training instructors and in roll calls, such that members of even large flights become very familiar with one another by name. All flights were assessed at the same point in their training: after 6 weeks of training together. The study was a round-robin design, in that each of the 809 participants acted as both a nominator ("judge") and a potential nominee ("target") in the peer nomination process.

Materials

Each participant was administered a computerized battery of measures, which included the self- and peer report versions of the Multisource Assessment of Personality Pathology (MAPP). Participants were first presented with a list of all other members of the flight and instructed as follows: "Please rate how well you know each person." This item was phrased as a rating-type item, such that participants were required to rate each group member, rather than nominating only the most relevant member(s). This item was rated using a 4-point rating scale ranging from 0 (*not at all*) to 3 (*very well*). Responses to the item were used to construct an affiliation matrix as described below.

The self- and peer-report versions of the MAPP each consist of 106 items, 81 of which are lay translations of the 10 DSM-IV personality disorder criteria. These personality disorder items were constructed by translating the DSM-IV criterion sets for PDs into lay language; resulting items were then reviewed and revised by expert consultants, including a member of the DSM-IV Personality Disorders Workgroup. Twenty-three filler items were also included in these measures, based on additional, mostly positive, characteristics, such as "trustworthy and reliable" or "agreeable and cooperative." The self-report and peer-report versions of items were otherwise identical, with only the relevant questions differing.

The peer-nomination procedure was a round-robin design in which every individual in the group had the opportunity to report on every other member of the group. Items were presented to participants in a quasi-random order. For each item, the participant was shown a list of all members of his or her group and asked to nominate at least one member of the flight who exhibited the characteristic in question. For each nomination, the participant assigned the nominee a score (1, 2, or 3), indicating that the nominee "sometimes," "often," or "always" displays the characteristic. Individuals who were not nominated for an item were tacitly given a score of 0. Peer-report scales, based on the DSM-IV criteria sets, were

calculated by averaging the scores received for the items in each scale, resulting in a dimensional scale ranging from 0 to 3. The scores assigned by each judge on each scale were kept separate for each target, such that in a flight with *N* members, each person received (*N* – 1) peer-report scores on each diagnostic scale.

Although scores by individual judges of targets were kept separate for most analyses, in some instances it was useful to combine reports of a target across all judges in order to conduct target-level analyses. In these cases, aggregate peer scores were constructed for each target by taking the mean of all judges'; reports for each of the diagnostic scales. Each target, therefore, had 10 aggregate peer scales, ranging from 0 to 3, which corresponded to the 10 peer diagnostic scales.

Following the peer-report section, all participants completed a self-report version of the same items. Participants were presented with the items in the same order and asked, "What do you think you are really like on this characteristic?" Participants responded using a 4-point scale: 0 (*never this way*), 1 (*sometimes this way*), 2 (*usually this way*), and 3 (*always this way*). For each personality disorder, the scores for the relevant criteria were averaged to form a dimensional measure of personality disorder ranging from 0 to 3.

We reported (Thomas, Turkheimer, & Oltmanns, 2003) on the psychometric properties and factor structure of the MAPP in a large military sample (*N* = 2,111 Air Force recruits, of which the present study uses a subsample). The interjudge reliability for peer reports on the MAPP items (i.e., the median coefficient alpha across groups, calculated across each of the judges for each PD feature) was .74 in the larger Air Force sample, with values ranging from .90 to .19 (although only three items had values below .5). Factor analysis of the peer-report items also demonstrated high correspondence (congruence coefficients ranged from .87 to .97) with factor patterns of widely used self-report models of PDs (Thomas et al., 2003).

Procedure

Two or three flights at a time were brought to a central testing center at Lackland Air Force Base. Each participant was seated at a separate computer terminal, where he or she gave written informed consent to participate in the study. After giving consent and before being administered the MAPP measures, each completed a computerized tutorial on how to select items by pointing and clicking using a mouse. The battery took an average of 2 hours to complete. During this time, participants were instructed not to talk to one another and to raise their hands if they encountered a problem or question. Dividers between workstations prevented participants from seeing the computer screens of those around them.

Data Analysis

Social network analysis—One of the central uses of social network analysis is finding subgroups of people who are more closely associated with one another. A variety of methods can be used (see Wasserman & Faust, 1994, for a full review). Depending on the method used, the subgroups can range from small, tight-knit cliques of close friends to large partitions indicating greater closeness relative to the surrounding network. For the present

study, we partitioned networks into groups using cluster analysis (e.g., Moody, 2001). We chose this methodology for several reasons. First, partitioning the network assigns each participant to a single group, ensuring that all participants will be included in the analyses. Moreover, cluster analysis is familiar to most social scientists and does not require dedicated social network analysis software, thue making this methodology more readily accessible.

For each flight, an adjacency matrix was constructed based on each participant's knowing score of each other individual. For a flight consisting of N participants, this consisted of an $N \times N$ directed matrix of how well (on a scale of 0 to 3) he or she knows each other individual. The adjacency matrices were transformed to distance matrices by subtracting each knowing score from 3, such that a knowing score of 3 was represented by a distance of 0, etc. In other words, for the purposes of analysis, the matrix was transformed to represent the perceived *distance* rather than the perceived closeness between each pair of participant.

This distance matrix was then submitted to the SAS Statistical Software Version 8.02 MODECLUS procedure (Method = 1) with a variable-radius kernel. This nonparametric procedure conducts cluster analysis by starting with each node in a separate cluster and joining neighbors together to form higher-density clusters (SAS Institute, 1999). The size of the cluster is determined partly by one or more specified smoothing parameters. In this case, the parameter specified, k, was the minimum number of individuals allowable in any cluster. To examine the effects of a greater number of subgroups, two separate sets of partitions were retained from each flight: one containing exactly two subgroups and one containing the maximum number of viable subgroups. First, we divided each flight into two groups ("two-partition clusters") by finding the smallest value of k that yielded a total of two clusters. Second, we divided each flight into the maximum number of groups for which each group contained at least three members ("maximal partition clusters"). This was done by retaining clusters from the largest value of k in which no cluster contained fewer than three members. In no case did increasing the number of clusters beyond this point result in a greater number of members in a cluster. The number of maximum clusters found for each flight is reported in Table 1.

Mixed model analysis—In order to evaluate the effect of shared social network subgroup membership on informant report, judge consensus and the effects of self-report were calculated separately within group and between groups. Separate analyses were conducted for shared two-partition cluster membership and shared maximal-partition cluster membership. Although there were 809 participants, there was a total of 31,314 judge-by-target dyads. In order to analyze multiple flights accurately, each containing judges reporting on multiple targets, and targets evaluated by multiple judges, it is necessary to model the data using an application of generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). Generalizability theory has been widely used to partition the variance in studies of interpersonal perception and judgment (e.g., Kenny, 1994; Malloy & Kenny, 1986; Shrout, 1993; Shrout & Fliess, 1979). The theory specifies that a report of a trait for a given target by a given judge varies from the grand mean based on several factors. These include the deviation of the judge overall (e.g., a tendency to rate more harshly than other judges), the deviation of the target overall (e.g., a tendency for all judges to identify the trait in the target), and specific interaction effects between the judge and target (e.g., a judge has

a particular dislike of a specific target). If, across multiple flights, multiple judges report on multiple targets, we can use generalizability theory to estimate the separate effect of flight membership, effect of judges, effect of targets, effect of specific judge-target interactions, and the residual error.

More formally, Equation 1, adapted from Shrout (1993) can be used to conceptualize the score given by Judge j to Target t:

$$X_{tj} = \mu + \alpha_f + \beta_t + \gamma_{tj} + \delta_{tj} + e_{tj} \text{ where:} \quad \text{(1)}$$

X_{tj} = The specific score given by Judge j to Target t

μ = The grand mean of all judges' reports of all targets on *d*.

α_f = The expected effect of Flight f

β_t = The expected effect of Target t

γ_j = The expected effect of Judge j

δ_{tj} = The specific impression of Target t by Judge j.

e_{tj} = Random error effects.

In other words, across all judges and targets, there is a grand mean for the score given by judges. Flights have some tendency to deviate from the grand mean, represented by α_f. Overall ratings of target t have some tendency, in general, to deviate from the mean of the flight, represented by β_t. Judge j has some tendency, in general, to deviate from the mean of the flight, represented by γ_j. In addition to these general effects, Judge j has a specific, idiosyncratic impression of Target t, represented by δ_{tj}. (In cross-sectional assessments, δ_{tj} is generally not distinguishable from random error effects, represented by e_{tj}, and in our calculations was subsumed within that term.) The combination of all of these effects leads to a specific report of Target t by Judge j on a given occasion (see Singer, 1998; Snijders & Bosker, 1994).

Based on these principles, linear mixed models were constructed for the data, using the MIXED procedure for SAS software. One limitation in using round-robin design to assess personality perceptions is nonindependence in the data, which Kenny's Social Relations Model describes as stemming from two sources (Kenny, 1994, 1996). First, an individual's target effect may correlate with his or her judge effect. That is, an individual who is seen by others as more narcissistic may see others as more narcissistic. Second, judges' and targets' dyadic impressions of one another may correlate. That is, if I see you as especially narcissistic, you may also see me as particularly narcissistic. These nonindependence correlations are difficult to model,[1] but tend to be quite small (Kenny, 1994).

---

[1] In order to more accurately estimate error variance due to nonindependence, one can also add a repeated measure component to the model, accounting for the dyadic correlation. However, exploratory analyses suggested that this component accounted for an extremely small amount of variance, with no change to the overall results, while vastly increasing the complexity of the model. In the interest of parsimony, the results reported here do not include this dyadic component. We thank an anonymous reviewer for assistance with the issue of nonindependence in our data.

Our analysis modeled both judge and target as random variables in linear mixed models. The basic equation specified for these models was: $y_{tj} = \mu + \alpha_{f} + \beta_{t(f)} + \gamma_{j(f)}$, where y_{tj} is the predicted value of a judgment, α_f, $\beta_{t(f)}$, and $\gamma_{j(f)}$ are all random effects variables, and $\beta_{t(f)}$ and $\gamma_{j(f)}$ are each nested within α_f. For each diagnostic category of personality disorder, this basic model was used to estimate the variance components of flight, judge, target, and residual (error) variance, using restricted maximum likelihood estimates. These variances were then used to calculate rater consensus, using the formula ICC $(2,1) = [\sigma^2(\beta)]/[\sigma^2(\beta) + \sigma^2(\gamma) + \sigma^2(e)]$. ICC(2,1) represents the average reliability of a single judge. Note that this calculation of consensus does not include γ_j as part of the numerator, which assumes that judges differ in their interpretation and use of scales but results in smaller ICC values (Hönekopp, 2006). We also calculated judge consensus using the formula ICC $(2,k) = [\sigma^2(\beta)]/[\sigma^2(\beta) + [\sigma^2(\gamma)/k] + [\sigma^2(e)/k]]$, where $k = (N - 1)$ and $N =$ the number of judges in the analysis, calculated by taking the harmonic mean for the number of judges across each flight or partition. ICC(2,k) provides an estimate of the reliability for all judges in a group.

In order to calculate self-peer agreement, we aggregated the peer-report data, as described above, finding each target's mean peer scores for all judges within a flight, and separately by shared two-partition and maximal-partition membership. We then calculated Pearson correlation coefficients between self-report and aggregated peer report for each group.

Results

Descriptive Statistics

As described above, acquaintance between each dyad was measured by self-report on a 4-point scale ranging from 0 to 3. The mean rating, across all dyads, for how well a judge knew a given target was 1.04 (*SD* = 1.02, 95% *CI* = 1.03–1.05). The reported acquaintance when both judge and target were in the same subgroup (ingroup) was compared with that of dyads for which the judge and target were in different subgroups (outgroup). For two-partition clusters, mean in-group knowing level (dyad *N* = 16,205) was 1.47 (*SD* = 1.00, 95% *CI* = 1.45–1.47), whereas for outgroup dyads (dyad *N* = 15,109), mean knowing was 0.58 (*SD* = 0.81, 95% *CI* = 0.57–0.59), an effect size of *d* = 0.97. For maximal-partition clusters, mean ingroup knowing level (dyad *N* = 7,079) was 1.73 (*SD* = 0.96, 95% *CI* = 1.70–1.75), whereas for outgroup dyads (dyad *N* = 24,235), mean knowing was 0.84 (*SD* = 0.94, 95% *CI* = 0.83–0.85), an effect size of *d* = 0.93.

Each of the 10 self-reported personality disorder scales ranged from a minimum possible value of 0 to a maximum possible value of 3. Mean scores ranged from 0.24 (antisocial) to 0.71 (OCPD). The means and standard deviations of each of the 10 scales is described in Table 2. The individual dyadic judgments of any given judge for any target also had a possible range from 0 to 3 on each of the 10 scales. When peer-report scores were examined across all participants, without partitioning the networks, mean scores ranged from 0.09 (avoidant) to 0.13 (narcissistic), as seen in Table 2.

Differences in mean scores were compared based on whether the target and judge shared a mutual cluster, for both two partitions per flight, and for the maximum number of viable partitions per flight. In both cases, ingroup judges assigned significantly higher scores,

although the effect size of the change increase was small and varied by diagnostic scale, as seen in Table 2. The effect of subgroup membership across all PD diagnostic categories was larger for two-partition clusters (Median d = .32) than maximal-partition clusters (Median d = .22).

Aggregated peer-report scores correlated moderately between in-group and outgroup judges. For two partitions, ingroup-outgroup correlations ranged from 0.39 (paranoid) to 0.61 (narcissistic), with a median correlation of 0.51. Maximal subgroup correlations were somewhat higher, ranging from 0.51 (schizoid) to 0.68 (histrionic), with a median correlation of 0.63.

Judge Consensus

To investigate judge consensus (i.e., how well any two judges agreed across targets), compact mixed linear models for each of the 10 diagnostic scales were constructed. The intraclass correlation, ICC(2,1), which describes the average consensus for a single judge, was calculated as described above. Intraclass correlation varied by diagnostic category, ranging from 0.13 (paranoid and schizoid) to 0.23 (narcissistic) with an overall median of 0.17 (see Table 3). We also estimated the consensus of all judges as a whole by calculating ICC(2,k), where k was defined as the harmonic mean of the number of raters minus 1. These values ranged from 0.84 (schizoid) to 0.92 (narcissistic) with an overall median of 0.88 (see Table 3).

After computing the consensus across all judges, the same process was repeated for networks partitioned into two subgroups and into the maximum number of subgroups. Consensus was calculated separately for ingroup and outgroup judges. In all cases, ingroup judges had higher levels of agreement than outgroup judges. In addition, ingroup judgments had higher levels of agreement than did the undifferentiated group (i.e., irrespective of partitions) for all diagnostic categories. As seen in Table 3, the improvement in ICC(2,1) based on two partitions ranged from 0.03 (antisocial PD) to 0.09 (paranoid PD), with a median improvement of 0.08 across all scales. The intraclass correlation for multiple judges showed improvements within partition that ranged from 0.03 (antisocial) to 0.13 (paranoid) with a median improvement of 0.09.

Improvement in ICC (2,1) for the maximum number of partitions ranged from 0.01 (schizoid PD) to 0.12 (dependent PD), with a median improvement of 0.06 across all PD scales. However, the smaller number of ingroup judges (harmonic mean = 8) compared to out-group judges (harmonic mean = 28) limited the ingroup benefits to ICC(2,k). Outgroup judges demonstrated higher consensus for all PD scales, ranging from 0.13 (dependent) to 0.28 (schizoid), with a median difference of 0.18.

Self-Peer Agreement

Self-peer agreement was calculated between self-report and aggregated peer report. Across diagnostic categories for all judges and targets irrespective of subgroup, the median self-peer correlation was 0.19. The self-peer correlation for each diagnostic category ranged from 0.14 (Narcissistic) to 0.30 (Avoidant) and appears in Table 4.

After calculating self-peer agreement for all participants, judges'; ratings of targets were aggregated on the basis of whether or not they were in a mutual subgroup, with separate analyses performed for two-partition and maximal-partition clusters. For each partition type, two sets of analyses were performed: one for targets and judges in the same subgroup and one for those who were not. Self-peer agreement for each of the 10 personality disorder scales was calculated separately for ingroup and outgroup judge-target pairings.

Partitioning the networks into two clusters demonstrated consistent improvements in self-peer correspondence ingroup relative to outgroup (see Table 4). However, compared with the full, undifferentiated group (i.e., all dyads), the ingroup self-peer agreement showed little improvement. That is, partitioning networks into two clusters did not identify judges who agreed more highly with self-report, but *did* identify those with worse self-peer agreement.

Partitioning the flights into the maximum number of viable clusters resulted in less consistent self-peer improvement than did two clusters. Several of the diagnostic categories (schizoid, schizotypal, narcissistic, avoidant, and OCPD) had higher outgroup self-peer agreement than ingroup. However, despite the less consistent improvement, the ingroup self-peer agreement for four of the diagnostic categories (paranoid, antisocial, histrionic, and dependent) was higher in maximal-partition clusters than in two-partition clusters.

Discussion

In this study, we applied social network analysis techniques to the assessment of self- and peer perceptions of personality pathology. The goal of the study was to determine whether these techniques, particularly those used to find cohesive subgroups within a network, could help us understand the limitations of peer reports obtained from large groups. The overall findings of this study suggest that there are, indeed, identifiable social network structures that can improve our understanding of peer perceptions. Specifically, ingroup peer judgments had higher mean scores, were more reliable, and, to some extent, corresponded more highly with self-report.

Mean Peer Report Scores

The mean values of ingroup evaluations were compared with out-group. The results of these comparisons, in Table 2, indicate that there is a highly significant tendency for ingroup peer-report scores to be higher than outgroup, with small to medium effect sizes. The largest effects were found for two-partition clusters; ingroup peer reports were on average more than twice as large as outgroup, with a median effect size of $d = 0.32$. Effects were smaller for maximal-partition clusters, with a median effect size of $d = 0.22$.

Due to the nomination procedure used to obtain peer reports, the increased ingroup mean scores represent both a greater number of nominations received and a higher average rating on those nominations. Although participants were free to nominate any member of the flight for each item, they were apparently more likely to nominate members of their own subgroups. This finding underscores the fact that shared cluster membership, particularly for two partitions, does not necessarily represent close friendship but, rather, overlapping groups of friends and mutual acquaintances within the social network. This result suggests

that individuals are more likely to nominate those within these larger acquaintance groups, perhaps because they had greater information on which to base their judgments.

Judge Consensus

As predicted, ingroup judge consensus was increased relative to out-group. As described above, for each diagnostic category the judge variance and residual variance of compact, mixed, linear models were used to calculate the intraclass correlation of judges. These results, across all participants, are summarized in Table 3. Consistent with cognitive social network research (e.g., Carley, 1986; Freeman, Romney, & Freeman, 1987), ingroup consensus was considerably higher than outgroup. The median ICC(2,1) across all dyads was 0.19, whereas for ingroup dyads, median ICC was 0.29. Similar improvements were seen in ICC (2,k) for two-partition subgroups (al-though not for the maximal number of clusters).

We suggest that the increased ingroup judge consensus is a reflection of interpersonal factors detailed in Kenny's (1994) Weighted Average Model (WAM), recently reparameterized as the PERSON model (Kenny, 2004). Kenny's model is a theoretical model to explain the sources of variance in person perception. It is made up of interrelated parameters, with relationships among these factors mediating acquaintance effects and leading to an overall consensus level. The clusters analyzed here can be seen as a surrogate measure for some of these factors, helping to explain the increased ingroup consensus. Four of the parameters from the WAM are particularly relevant to the present study: Acquaintance, Overlap, Similar Meaning Systems, and Communication.

First, the increased consensus is likely associated with the increased *acquaintance* (the amount of information the perceiver has about the target), as reflected by increased ingroup knowing scores. Numerous studies have demonstrated increased consensus (e.g., Funder & Colvin, 1988) as a function of increased acquaintance. However, although maximal-partition clusters identified subgroups that were more highly acquainted than two-partition subgroups (Mean ingroup knowing = 1.74 vs. 1.47, respectively), consensus was more improved in the two-partition subgroups. Consistent with Kenny (1994), this suggests that other factors, such as overlap in observations, similar meaning systems by judges, and communication among judges and may play a larger role in consensus.

Overlap is the extent to which two judges observe the same set of behaviors by the target (Kenny, 2004). Although two perceivers might have similar amounts of information about a target (i.e., acquaintance), if they know the individual from different contexts, their judgments may be based on observation of very different behaviors. Prior social network studies have found that judges within a cohesive subgroup are likely to be exposed to the same sort of information (i.e., overlap), leading them to make similar judgments. For example, Freeman and colleagues (1987) found that participants within cohesive subgroups were exposed to the same sorts of information about other members of the network, leading to increased consensus even when these judgments were wrong.

Similar meaning systems refers to the extent to which two observers, observing the same behavior, impute the same meaning to the behavior (Kenny, 1994). Similar meaning systems

suggest that in order for judges to have high consensus, they must interpret targets'; behavior in the same way. There is some evidence that individuals within cohesive subgroups may have more similar meaning systems than those between subgroups. This theory of structural balance is supported by a long tradition in both social psychology (e.g., Festinger, 1957) and social network analysis (e.g., Johnsen, 1986). In addition, social psychological studies of group polarization (e.g., Moscovici & Zavalloni, 1969) also indicate that small groups of individuals with similar attitudes can become polarized, such that the strength of the attitudes become enhanced for all group members. It therefore seems likely that the increased consensus found within subgroups is at least partly due to the effects of increased similar meaning systems among group members.

Finally, *communication* is an estimate of judges'; discussion of their impressions of targets, such that judges who discuss their opinions of a target may converge in their ratings (Kenny, 1994). Communication among judges is particularly complicated to model and has largely been neglected in the literature (Kenny, 2004). However, the social network tradition, which Pattison (1994) describes as "Information Bias," considers network connections as conduits for information that would lead to greater communication among judges within a cohesive subgroup. For example, Carley (1986) noted that members of a network gained information based on their pattern of connections with others, enhancing consensus within cohesive subgroups.

To summarize, comparing ingroup and outgroup judge consensus found substantially higher ingroup consensus. The magnitude of this improvement varied slightly by method of subgrouping, with two clusters providing greater improvement than numerous smaller clusters. When multiple rater consensus (ICC [2,k]) was calculated, improvements were found for two clusters, but not for multiple small clusters, due to the relatively small number of judges making within-group ratings.

Self-Peer Agreement

The second question of interest in this study was the effect of mutual subgroup membership on self-peer agreement. We hypothesized that ingroup self-peer agreement would be higher than outgroup, which was partially supported, although effects were small. As described above, we calculated the correlation between self-report and peer report aggregated across all raters or by subgroup for two-partition and maximal-partition clusters (see Table 4).

For two-partition clusters, ingroup assessments demonstrated consistently higher self-peer agreement than outgroup. For maximal-partition clusters, ingroup self-peer agreement was also higher, on average, although these findings were less consistent than for two-partition clusters. These findings are consistent with previous studies of self-peer agreement as a function of acquaintance. Numerous cross-sectional (e.g., Funder & Colvin, 1988) and longitudinal (e.g., Paulhus & Bruce, 1992) studies have found that greater acquaintance is associated with increased self-peer agreement.

One explanation for the present finding is that ingroup peers had greater opportunity to interact with the target and to learn the target's own self-perceptions. Peers might assimilate this information into their own impressions of the target, making ratings that were more

highly congruent with the target's self-report, regardless of the validity of the judgments. A second explanation is that differences in self-peer agreement reflect the improved reliability of ingroup judges. This is supported by the fact that dividing a flight into numerous smaller clusters provided less consistent improvements than did identifying two large clusters. As noted earlier, creating numerous clusters relegates judges who know the target moderately well to an outgroup status, which appears to inflate the outgroup self-peer agreement. In other words, creating two partitions primarily removes those judges who do not know a target well, whereas multiple partitions may set the threshold too high.

Limitations and Future Research

Several limitations of the present study should be noted. First, it is unclear how well these findings will generalize to the larger population. We chose to study military trainees because the setting provided some significant benefits to the investigation of interpersonal perception. In particular, recruits did not know one another prior to training and were randomly assigned to flights, controlling for the length of their acquaintance with one another (Oltmanns & Turkheimer, 2006). Although the participants were representative of the general populace in many ways, their social situation was unique in that much of their time and dealings with others were constrained by the requirements of training. Further research to generalize these findings should attempt to extricate the effects of the military environment from subgroup formation.

That said, the constraints of military life should not be interpreted as limiting the particular findings of this study. Although the participants' social environment was affected by some unusual restrictions, it is likely that factors associated with subgroup formation in any type of social network will be difficult to generalize. The social forces in a college dorm are very different from those in a workplace, which are, in turn, different from those at a country club. More important to the findings presented here are the implications for self-peer agreement and judge-judge consensus associated with mutual subgroup membership.

A second area for improvement on the current study is an enhanced measurement of the social networks themselves. Acquaintance was assessed based on a single question about how well the judge knew each target. In a more complete social network study, acquaintance might be assessed using more specific questions. In addition, the network structure might be assessed in ways other than self-report. Asking participants to identify friendships between other dyads, or observations by an outside party, might yield a different picture of the network than that derived from self-report data alone (e.g., Bernard et al., 1984).

Third, the present study focused on cluster analysis to determine cohesive subgroups. Cluster analysis techniques are useful tools for social network analysis, particularly because they are available in most statistical software packages without the need to obtain specialized SNA software. However, one of the limitations of cluster analysis is its reliance on local maxima for density estimation. Dedicated SNA software like UCInet (Borgatti, Everett, & Freeman, 2002) can apply more sophisticated methods (e.g., TABU algorithms; Pan, Chu, & Lu, 2000) to determine cohesive subgroups. Analysis of these methods using the techniques described in the current study suggests that improvements to reliability and

self-peer agreement are nearly identical to those gained from cluster analysis (Clifton, 2005).

Finally, the present study examined two subjective measures of personality, self-report and peer report. Although we found evidence for both increased ingroup judge consensus and increased ingroup self-other agreement, neither measure should be interpreted as indicating greater validity of judgments. Rather, we see these findings as a starting point for future research examining more objective indicators of functioning. For example, the authors have previously examined the predictive ability of self- and peer reports of personality disorders in predicting early discharge from the military (Fiedler et al., 2004), impaired social functioning (Miller, Pilkonis, & Clifton, 2005; Oltmanns, Melley, & Turkheimer, 2002), and performance in laboratory tasks (South, Oltmanns, & Turkheimer, 2003). Future research should examine peers'; accuracy in predicting similar objective indicators of behavior as a function of their social network position and characteristics.

Implications and Conclusions

The findings of the present study indicate that when informant reports are partitioned on the basis of social network analysis, ingroup informants provide different information from, and are more reliable than, outgroup informants. The benefits seen in dividing networks into two partitions were more consistent and generally larger than those obtained from a greater number of small subgroups. Dividing a network into two partitions seems to be adequate to identify knowledgeable judges, whereas increasing the number of partitions may inadvertently exclude desirable informants from the analysis.

The implications of these results are twofold. First, our findings suggest that social network analysis is useful in identifying more desirable informants. When assessing large groups of informants, including information from all informants is optimal. In practice, however, constraints on time or resources may limit the number of informants that can be obtained for each target. In cases in which researchers must select a limited number of informants, SNA can inform the choice of more reliable informants. For example, in the present study, selecting five ingroup judges would, on average, result in an ICC of 0.64, compared with a reliability of 0.51 for the same number of judges selected at random from the whole group.

Second, our findings suggest that there are very real differences in the information provided by ingroup and outgroup informants. When informants are divided into two partitions, judgments by ingroup and outgroup judges correlate only moderately (median $r = 0.51$). Moreover, ingroup judges report a significantly greater amount of pathology than outgroup judges (see Table 2), which could result in different estimates of prevalence or severity depending on the judges assessed. Without an external measure of validity, the relative utility of information from ingroup and outgroup sources remains unknown. However, it is likely that ingroup and outgroup judgments are based on different observations of behavior and on different channels of communication, both target-judge and judge-judge. Our findings suggest that these factors, often neglected in informant studies, should be more carefully assessed. We encourage further investigation and application of network analysis techniques in the assessment of normal and maladaptive personality.

Acknowledgments

References

Bernard HR, Killworth PD, Kronenfeld D, Sailer L. The problem of informant accuracy: The validity of retrospective data. Annual Review of Anthropology. 1984; 13:495–517.

Borgatti, SP.; Everett, MG.; Freeman, LC. Ucinet 6 for Windows. Harvard, MA: Analytic Technologies; 2002. Computer software and manual

Breiger RL, Ennis JG. Personae and social roles: The network structure of personality types in small groups. Social Psychology Quarterly. 1979; 42(3):262–270.

Carley KM. An approach for relating social structure to cognitive structure. Journal of Mathematical Sociology. 1986; 12:137–189.

Clifton, AD. Unpublished doctoral dissertation. University of Virginia; Charlottesville: 2005. Social network analysis of self and peer perceptions of personality pathology.

Clifton A, Turkheimer E, Oltmanns TF. Contrasting perspectives on personality problems: Descriptions from the self and others. Personality and Individual Differences. 2004; 36:1499–1514.

Clifton A, Turkheimer E, Oltmanns TF. Self and peer perspectives on pathological personality traits and interpersonal problems. Psychological Assessment. 2005; 17:123–131. [PubMed: 16029099]

Cronbach, L.; Gleser, GC.; Nanda, H.; Rajaratnam, N. The dependability of behavioral measurements: Theory of generalizability of scores and profiles. New York: Wiley; 1972.

Festinger, L. A theory of cognitive dissonance. Stanford, CA: Stanford University Press; 1957.

Fiedler ER, Oltmanns TF, Turkheimer E. Traits associated with personality disorders and adjustment to military life: Predictive validity of self and peer reports. Military Medicine. 2004; 169:207–211. [PubMed: 15080240]

Freeman LC, Romney AK, Freeman SC. Cognitive structure and informant accuracy. American Anthropologist. 1987; 89:310–325.

Funder DC, Colvin CR. Friends and strangers: Acquaintanceship, agreement, and the accuracy of personality judgment. Journal of Personality & Social Psychology. 1988; 55:149–158. [PubMed: 3418488]

Funder DC, West SG. Consensus, self-other agreement, and accuracy in personality judgment: An introduction. Journal of Personality. 1993; 61:457–476. [PubMed: 8151499]

Hönekopp J. Once more: Is beauty in the eye of the beholder? Relative contributions of private and shared taste to judgments of facial attractiveness. Journal of Experimental Psychology. 2006; 32:199–209. [PubMed: 16634665]

Johnsen EC. Structure and process: Agreement models for friendship formation. Social Networks. 1986; 8:257–306.

Kanfer A, Tanaka JS. Unraveling the web of personality judgments: The influence of social networks on personality assessment. Journal of Personality. 1993; 61:711–738. [PubMed: 8151505]

Kenny, DA. Interpersonal perception: A social relations analysis. New York: The Guilford Press; 1994.

Kenny DA. Models of nonindependence in dyadic research. Journal of Social and Personal Relationships. 1996; 13:279–294.

Kenny DA. PERSON: A general model of interpersonal perception. Personality and Social Psychology Review. 2004; 8:265–280. [PubMed: 15454349]

Kenny DA, Albright L, Malloy TE, Kashy DA. Consensus in interpersonal perception: Acquaintance and the Big Five. Psychological Bulletin. 1994; 116:245–258. [PubMed: 7972592]

Kenny DA, Kashy DA. Enhanced co-orientation in the perception of friends: A social relations analysis. Journal of Personality & Social Psychology. 1994; 67:1024–1033. [PubMed: 7815299]

Klonsky ED, Oltmanns TF, Turkheimer E. Informant reports of personality disorder: Relation to self-report, and future research directions. Clinical Psychology: Science and Practice. 2002; 9:300–311.

Malloy TE, Kenny DA. The social relations model: An integrative methodology for personality research. Journal of Personality. 1986; 54:199–225.

McCrae RR, Costa PT. Validation of the five-factor model of personality across instruments and observers. Journal of Personality and Social Psychology. 1987; 52:81–90. [PubMed: 3820081]

Michaelson A, Contractor NS. Structural position and perceived similarity. Social Psychology Quarterly. 1992; 55:300–310.

Miller JD, Pilkonis PA, Clifton A. Self and other-reports of traits from the Five-Factor Model: Relations to personality disorder. Journal of Personality Disorders. 2005; 19:400–419. [PubMed: 16178682]

Moody J. Peer influence groups: Identifying dense clusters in large networks. Social Networks. 2001; 23:261–283.

Moscovici S, Zavalloni M. The group as a polarizer of attitudes. Journal of Personality and Social Psychology. 1969; 12:125–135.

Oltmanns TF, Melley AH, Turkheimer E. Impaired social functioning and symptoms of personality disorders in a non-clinical population. Journal of Personality Disorders. 2002; 16:438–453.

Oltmanns, TF.; Turkheimer, E. Perceptions of self and others regarding pathological personality traits. In: Krueger, RF.; Tackett, J., editors. Personality and psychopathology. New York: Guilford; 2006. p. 71-111.

Pan, JS.; Chu, SC.; Lu, ZM. Cluster generation using tabu search based maximum descent algorithm. In: Ebecken, N.; Brebbia, CA., editors. Data Mining II/International Conference on Data Mining. Cambridge, UK: WIT Press; 2000. p. 525-534.

Park B, Judd CM. Agreement on initial impressions: Differences due to perceivers, trait dimensions, and target behaviors. Journal of Personality & Social Psychology. 1989; 56:493–505.

Pattison, P. Social cognition in context: Some applications of social network analysis. In: Wasserman, S.; Galaskiewicz, J., editors. Advances in social network analysis: Research in the social and behavioral sciences. Thousand Oaks, CA: Sage; 1994. p. 79-109.

Paulhus DL, Bruce MN. The effect of acquaintanceship on the validity of personality impressions: A longitudinal study. Journal of Personality and Social Psychology. 1992; 63:816–824.

SAS Institute Inc.. SAS OnlineDoc, Version 8. Cary, NC: SAS Institute Inc; 1999. Computer manualRetrieved October 23, 2002, from http://v8doc.sas.com/sashtml/

Shrout PE. Analyzing consensus in personality judgments: A variance components approach. Journal of Personality. 1993; 61:769–788. [PubMed: 8151506]

Shrout PE, Fliess JL. Intraclass correlation: Uses in assessing rater reliability. Psychological Bulletin. 1979; 86:420–428. [PubMed: 18839484]

Singer JD. Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. Journal of Educational and Behavioral Statistics. 1998; 24:323–355.

Snijders TAB, Bosker RJ. Modeled variance in two-level models. Sociological Methods and Research. 1994; 22:342–363.

South SC, Oltmanns TF, Turkheimer E. Personality and the derogation of others: Descriptions based on self and peer report. Journal of Research in Personality. 2003; 37:16–33.

Tajfel, H., editor. Differentiation between social groups. New York: Academic Press; 1978.

Thomas RC, Turkheimer E, Oltmanns TF. Factorial structure of pathological personality as evaluated by peers. Journal of Abnormal Psychology. 2003; 112:81–91. [PubMed: 12653416]

Wasserman, S.; Faust, K. Social network analysis: Methods and applications. New York: Cambridge University Press; 1994.

Table 1

Descriptive Statistics and Maximum Number of Clusters Obtained From Each Flight Group

Flight	N	N Male (%)	Density (Mean Acquaintance)	Number of Maximal Clusters
1	36	21 (58.3%)	0.97	5
2	33	20 (60.6%)	0.94	4
3	36	36 (100%)	0.88	6
4	35	35 (100%)	0.88	6
5	41	22 (53.6%)	1.01	6
6	38	21 (55.3%)	1.09	5
7	35	19 (54.3%)	1.15	5
8	37	19 (51.4%)	1.02	8
9	37	22 (59.5%)	1.14	6
10	34	21 (61.8%)	1.18	3
11	41	41 (100%)	1.07	4
12	41	41 (100%)	1.03	4
13	35	21 (60.0%)	1.18	3
14	35	17 (48.6%)	1.11	6
15	27	11 (40.7%)	1.17	4
16	39	16 (41.0%)	1.28	4
17	36	18 (50.0%)	0.98	4
18	42	42 (100%)	0.95	6
19	45	45 (100%)	1.07	7
20	52	17 (32.7%)	1.00	5
21	54	28 (51.9%)	0.98	6
Total	809	533 (65.9%)	N/A	107
Mean	38.5		1.05	5.1
SD	6.13		0.11	1.3

Table 2

Mean Dyadic MAPP Personality Disorder Scores for Self-Report and Peer Report Across All Judges and Partitioned by Clusters

Diagnostic Category	Self Report (N=809)		All Judges (N=31,314)		Two Clusters					Maximal Clusters					
					Outgroup (N = 15,109)		Ingroup (N=16,205)			Outgroup (N=24,235)		Ingroup (N=7,079)			
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Effect d	Mean	SD	Mean	SD	Effect d
Paranoid	0.49	0.46	0.10	0.25	0.05	0.17	0.14	0.29	0.34	0.08	0.23	0.15	0.30	0.24
Schizotypal	0.37	0.39	0.09	0.22	0.06	0.17	0.12	0.25	0.29	0.08	0.21	0.13	0.25	0.19
Schizoid	0.62	0.41	0.11	0.23	0.08	0.19	0.15	0.26	0.30	0.10	0.22	0.15	0.25	0.20
Antisocial	0.24	0.32	0.10	0.27	0.06	0.21	0.13	0.31	0.26	0.09	0.25	0.14	0.31	0.18
Borderline	0.26	0.34	0.09	0.21	0.05	0.15	0.13	0.25	0.36	0.08	0.19	0.14	0.26	0.27
Histrionic	0.37	0.36	0.11	0.26	0.07	0.19	0.15	0.31	0.34	0.10	0.24	0.16	0.33	0.25
Narcissistic	0.29	0.33	0.13	0.34	0.09	0.26	0.17	0.39	0.26	0.12	0.31	0.17	0.40	0.18
Avoidant	0.39	0.41	0.09	0.23	0.06	0.17	0.12	0.26	0.26	0.08	0.21	0.12	0.27	0.19
Dependent	0.25	0.33	0.09	0.25	0.05	0.17	0.13	0.30	0.33	0.08	0.22	0.15	0.32	0.25
OCPD	0.71	0.40	0.12	0.23	0.07	0.17	0.16	0.26	0.40	0.10	0.21	0.18	0.27	0.30

Note: All comparisons between Outgroup and Ingroup mean scores are significant: $F_{(1,31312)}$, $p<.0001$.

Table 3

Consensus (Intraclass Correlation) for All Judges, Outgroup Judges, and Ingroup Judges

| | ICC (2,1) | | | | | ICC (2, k) | | | | |
| | | Two Clusters | | Maximal Clusters | | | Two Clusters | | Maximal Clusters | |
Diagnostic Category	All Judges	Outgroup	Ingroup	Outgroup	Ingroup	All Judges (k = 37)	Outgroup (k = 17)	Ingroup (k = 18)	Outgroup (k = 27)	Ingroup (k = 7)
Paranoid	0.13	0.13	0.22	0.13	0.20	0.85	0.71	0.84	0.80	0.64
Schizoid	0.13	0.13	0.17	0.13	0.14	0.84	0.71	0.79	0.81	0.52
Schizotypal	0.19	0.18	0.27	0.18	0.23	0.89	0.78	0.87	0.86	0.67
Antisocial	0.22	0.26	0.29	0.22	0.25	0.91	0.85	0.88	0.89	0.70
Borderline	0.18	0.17	0.26	0.17	0.24	0.89	0.78	0.86	0.85	0.68
Histrionic	0.21	0.21	0.28	0.20	0.27	0.91	0.82	0.88	0.87	0.72
Narcissistic	0.23	0.24	0.30	0.23	0.28	0.92	0.84	0.89	0.89	0.72
Avoidant	0.16	0.15	0.24	0.15	0.21	0.87	0.74	0.85	0.83	0.65
Dependent	0.16	0.16	0.25	0.14	0.25	0.87	0.76	0.86	0.82	0.69
OCPD	0.14	0.14	0.20	0.15	0.19	0.85	0.72	0.82	0.82	0.62
Median	0.17	0.17	0.26	0.16	0.23	0.88	0.77	0.86	0.84	0.68

Table 4

MAPP Scales Correlation Between Self and Peer Reports by All Judges, Outgroup Judges, and Ingroup Judges

	All Judges	Two Partition Clusters		Maximal Partition Clusters	
		Outgroup	Ingroup	Outgroup	Ingroup
Paranoid	0.15	0.07	0.17	0.14	0.15
Schizoid	0.23	0.17	0.21	0.21	0.15
Schizotypal	0.28	0.16	0.28	0.27	0.24
Antisocial	0.19	0.14	0.19	0.17	0.21
Borderline	0.22	0.11	0.22	0.20	0.19
Histrionic	0.17	0.11	0.16	0.14	0.17
Narcissistic	0.14	0.14	0.13	0.15	0.11
Avoidant	0.30	0.18	0.29	0.29	0.25
Dependent	0.18	0.09	0.16	0.15	0.19
OCPD	0.14	0.08	0.15	0.14	0.11
Median	0.19	0.13	0.18	0.16	0.18