Published in final edited form as:

Assessment. 2015 December ; 22(6): 655–664. doi:10.1177/1073191114567942.

Psychometric Comparison of Self- and Informant-Reports of Personality

Thomas M. Olino¹ and Daniel N. Klein²

¹Temple University, Philadelphia, PA, USA

²Stony Brook University, Stony Brook, NY, USA

Abstract

Self-reports are the most relied on assessment method in psychology. In the area of personality, informant-reports are a reasonable alternative assessment strategy. However, agreement between self- and informant-reports of personality is only moderately good. A portion of the observed discrepancies between self- and informant-reports of personality may come from differences in psychometric measurement across raters. That is, it is unknown whether the constructs assessed via self- and informant-reports are psychometrically identical. We examined four key personality scales—Well-Being, Social Closeness, Stress Reaction, and Harm Avoidance—in male and female dyads who provided self- and informant-reports for their partner. Similarities in self- and informant-reports of personality were evaluated by testing measurement invariance. Overall, models supported configural, metric, and scalar invariance for each of the four personality dimensions. These results suggest that the same psychometric constructs are assessed via self- and informant-reports of these personality dimensions. Informant-reports can be used in studies to avoid biases from relying solely on self-reports.

Keywords

personality; measurement invariance; informant-reports; self-reports

Self-reports are the most relied on assessment strategy in the field of psychology. These measures are often inexpensive and much less burdensome on participants and researchers than alternative assessment strategies. However, there are numerous concerns about potential biases introduced when relying on self-reports. Furthermore, as many studies rely on participants to provide information about all constructs within the same study, there are concerns that mono-method approaches are overestimating the “true” associations reported in many studies (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003).

Reprints and permissions: sagepub.com/journalsPermissions.nav

Corresponding Author: Thomas M. Olino, Department of Psychology, Temple University, 1701 N. 13th St, Philadelphia, PA 19122, USA. thomas.olino@temple.edu.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Use of alternative assessment methods is common within subdisciplines of psychology. Neuropsychological and intelligence tests are used to assess cognitive function (Lezak, 2004). Interviewer-based instruments are often used in clinical assessments (Kessler et al., 2004; Kessler & Ustün, 2004). Behavioral observations and/or parent-reports are used in assessments of child temperament (De Los Reyes & Kazdin, 2005; Durbin, Hayden, Klein, & Olino, 2007; Fox, Henderson, Marshall, Nichols, & Ghera, 2005). The primary alternative in studies of adult personality is informant reports (although see Borkenau, Riemann, Angleitner, & Spinath, 2001; Vazire, 2006, for other approaches). However, important questions remain about whether the constructs assessed by self-reports are the same when assessed via alternative methods.

In the context of personality assessment, informant- and self-reports share a number of important features, including similarities in the administration method, required cognitive or functional level of the informants, and (often) scaling of the responses (see Ready & Clark, 2002, for an exception). However, there are a number of potential limitations of using informant reports. One concern is feasibility. Relying on an additional individual to provide information increases burden on the participants (and research team) and increases expense to conduct the research. Fortunately, previous research indicates that these additional burdens are generally not prohibitive in either moderately sized (Vazire, 2006) or even large samples (Clifton, Turkheimer, & Oltmanns, 2004, 2005).

An additional concern about the use of informant-reports is convergent validity with self-reports of personality. Meta-analytic results demonstrate substantial agreement between self- and informant-reports of adult personality (Connolly, Kavanagh, & Viswesvaran, 2007). The authors reported cross-sectional effect sizes of .46 for agreeableness, .56 for conscientiousness, .51 for emotional stability/neuroticism, .62 for extraversion, and .59 for openness to experience after adjusting for measurement error. While the magnitude of associations demonstrates good correspondence between raters, there is still much unique variance provided by both the targets and the informants. Moreover, Harkness, Tellegen, and Waller (1995) found that convergence between self- and informant-reports of personality differed across type of informant (e.g., mothers, fathers, peers) and dimensions of personality.

Agreement can also be assessed by examining mean-level differences between self- and informant-reports of personality. Surprisingly, few studies have reported these types of analyses. Some of these studies have focused on personality disorders (Riso, Klein, Anderson, Ouimette, & Lizardi, 1994; South, Oltmanns, Johnson, & Turkheimer, 2011) and others have emphasized normative personality dimensions (Flett, Besser, & Hewitt, 2005; McCrae et al., 2004; Miller, Pilkonis, & Clifton, 2005). Overall, these studies find lower levels of pathological personality using informant- than self-reports. However, mean level differences in normative personality traits were nonsignificant.

Importantly, self- and informant-reports of personality differentially predict a number of meaningful outcomes. A recent meta-analysis found that informant-reports of personality were better predictors of job performance than self-reports (Oh, Wang, & Mount, 2011). Self- and informant-reports of personality psychopathology have also demonstrated

differential predictive validity. For example, Klein (2003) found that both self- and informant-reports of personality disorders were associated with depressive symptoms and global functioning, but only informant-reports were associated with interpersonal functioning.

Discrepancies between self- and informant-reports can come from at least three plausible sources: self-report biases, access to different information, and psychometric differences across informants. Self-report biases can take many forms (Podsakoff et al., 2003). Respondents may have implicit theories about psychological constructs that influence responses on multiple items, which may bias sets of responses. Respondents may incorporate social desirability in their responses, such that they present themselves as more consistent with social norms than their actual attitudes or behavior. Alternatively, acute mood-state may influence ratings of personality (De Los Reyes & Kazdin, 2005; Durbin & Wilson, 2012; Hayden, Durbin, Klein, & Olino, 2010). For example, two recent studies reported that maternal personality and state depression influenced the associations between maternal-reports of their children's temperament and observer temperament in a laboratory assessment (Durbin & Wilson, 2012; Hayden et al., 2010). In addition, individuals providing self-reports and informants providing collateral-reports will have access to different information regarding the targets. For example, informants may see targets in only a limited number of contexts. Similarly, some personality measures assess traits over long periods of time including periods when the target and informant did not know each other or had little contact. Empirically, there is evidence that when the duration of familiarity is longer, there is greater correspondence between self- and informant-reports of personality (Connolly et al., 2007). However, even among individuals who have spent substantial time with one another in multiple contexts (e.g., spouses), members of the dyad have time apart and have private thoughts and feelings that are not accessible to the other. These characteristics are noted in several models of associations between self- and other-ratings (Funder, 2012; John & Robins, 1994; Vazire, 2010).

Finally, discrepancies between self- and informant-reports of personality could be due to differences in the psychometric construct assessed. That is, although the same items may be used to assess the underlying trait, the relationship between the underlying trait and the items may differ across informants. This could occur for a number of reasons and in a number of ways. For example, Vazire (2006) argues that self-perception may affect self-reports more strongly than informant-reports. Individuals' specific self-report responses may be positively influenced by general positive attitudes toward the self, and negatively by negative attitudes toward the self. This would lead to different rank-orderings of individuals when comparing self- and informant ratings, but might not affect mean levels of self- and informant-reported traits. In contrast, responses about others tend to be more consistently positive than self-reports. This should lead to more positive mean ratings based on informant-, compared to self-, report, but might not affect rank-orderings between individuals. Alternatively, differences in access to information about thoughts and behaviors may have an impact on the responses to items. Items concerning observable behaviors may be equally accessible for self- and informant-reports. However, items that assess internal states, preferences, and thoughts may only be accessible through self-reports. This could

produce differences in both rank-order and mean-levels of self- and informant-reported traits.

In each of these cases, the structure of the scale (or pattern of correlations between items) will differ for self- versus informant-report. The term *measurement invariance* (Vandenberg & Lance, 2000; Widaman, Ferrer, & Conger, 2010) refers to whether the measurement functioning of the target construct is comparable across different informants, groups, or occasions. In the absence of measurement invariance, it is not possible to determine whether observed differences reflect true differences in the construct and/or differences in measurement properties (Widaman et al., 2010).

There are multiple levels of measurement invariance that address different questions. The most basic requirement is that the same items are associated with the same construct for both informants (i.e., do the same items load on the same factors?). This is referred to as configural invariance. If the items assessing what are purportedly the same constructs differ across raters, the items hold different meanings for the informants. Next, it is important for the relationship between an item and the underlying construct be the same across informants. That is, the associations between the items and the underlying construct should be similar across informants. This is referred to as metric invariance. Finally, the probability of endorsing a specific item in the same manner should be the same across informants (Widaman et al., 2010). This is referred to as scalar invariance. When at least these three levels of measurement invariance are present in the data, scale scores reflect the same psychometric quantity and observed differences across characteristics can be attributed to the characteristic of interest, as opposed to psychometric factors. However, it remains untested whether self- and informant-reports of personality differ in their psychometric properties.

Measurement invariance of personality measures has been studied across genders (Laverdiere, Morin, & St-Hilaire, 2013; Marsh et al., 2010; Marsh, Nagengast, & Morin, 2013; Merz et al., 2013; Mor et al., 2008; Reise, Smith, & Furr, 2001; Rowinski, Ciecuch, & Oakland, 2014; Smith & Reise, 1998), developmental periods (Laverdiere et al., 2013; Marsh et al., 2010; Merz et al., 2013; Rowinski et al., 2014; Spence, Owens, & Goodyer, 2013), and cultures (Church et al., 2011; Johnson, Spinath, Krueger, Angleitner, & Riemann, 2008; Woo et al., 2014). However, whether self- and informant-reports of personality differ in their psychometric properties has not been previously examined.

The present article examines whether there are psychometric differences between self- and informant-reports of personality. More specifically, do self- and informant-reports of personality demonstrate measurement invariance? In the absence of previous data on the topic, we offer two tentative hypotheses. First, we hypothesize that constructs assessed using items describing more observable behaviors will be more consistently associated with the underlying construct across informants than items that describe individuals' personal preferences or cognitive style. Thus, there may be differences in factor loadings according to rater based on whether items are observable. Second, although some models of self-informant assessment argues that self-perception may influence one's ratings in either direction (e.g., Vazire, 2010), we speculate that, on average, informants will be more likely

to endorse socially undesirable traits more than the self. Thus, there may be differences in the thresholds at which individuals and their informants will endorse these items. As previous studies (Connolly et al., 2007) demonstrate that closeness of relationship moderates the association between self- and informant-reports, we sought to reduce variability on this dimension. Thus, we relied on self- and informant-reports of personality from a sample of married and/or cohabiting men and women.

Method

The data for this study come from the Stony Brook Temperament Study (Olino, Klein, Dyson, Rose, & Durbin, 2010), which is a longitudinal study of temperament and psychopathology beginning when children were 3 years old. The sample consisted of 559 families from a suburban community. Potential participants were identified using a commercial mailing list and screened by telephone. Families with a child between 3 and 4 years of age, who lived with an English-speaking biological parent and did not have significant medical conditions or developmental disabilities were included. Of eligible families, 66.4% entered the study. Families who agreed and declined to participate did not differ significantly on child sex and race/ethnicity and parental marital status, education, and employment status. Mothers were 36.0 years ($SD = 4.5$) and fathers were 38.3 years ($SD = 5.4$) old. Most child participants (86.9%) were White and came from middle class families, as measured by Hollingshead's Four Factor Index of Social Status ($M = 45.1$; $SD = 10.9$). Approximately half the mothers (54.7%) and fathers (45.7%) had at least a 4-year college degree. Most (96.5%) dyads were married or living together and 51.9% of the mothers worked outside the home part- or full-time.

Self-reports were collected from women and men using the brief form of the Multidimensional Personality Questionnaire (Patrick, Curtin, & Tellegen, 2002). This 155-item inventory assesses broad and narrow domains of personality. Items are rated using a dichotomous response format. The Multidimensional Personality Questionnaire assesses Positive Emotionality (PE), with narrow facets of Well-Being, Social Potency, Social Closeness, and Achievement; Negative Emotionality (NE), with narrow facets of Stress Reaction, Alienation, Aggression, and Control; and Constraint (CON), with narrow facets of Control, Harm Avoidance, and Traditionalism.

We also collected informant-reports from the other member of the couple regarding the target participant. That is, husbands served as the informant on their wife's personality and vice versa. To reduce burden while assessing content domains across the three broad domains, informants were only asked to complete a subset of the narrow-band scales. Consistent with previous studies of informant-reports of personality, informant questions were reworded in the third person (e.g., Bagby et al., 1998; Markon, Quilty, Bagby, & Krueger, 2013). The informant questionnaire included the Well-Being (14 items), Social Closeness (12 items), Stress Reaction (15 items), and Harm Avoidance (12 items) scales. These four scales were selected for two primary reasons. First, there is at least one scale for each superfactor, which provides coverage of the big three domains. Second, these scales were deemed to be most relevant to the broader project goals of relating personality and temperament to risk for internalizing psychopathology. Means, standard deviations, and

internal consistency estimates (α) for self- and informant-report reports are provided in Table 1. All internal consistency estimates are at least good, with the exception of self- and informant-reports of harm avoidance in women, which is adequate. Furthermore, internal consistency estimates and means are consistent with those reported in the community sample used to develop the brief form (Patrick et al., 2002).

Personality data were available on 482 women (473 self-reports; 395 informant-reports) and 476 men (397 self-reports; 466 informant-reports). Availability of self-reports by women was unrelated to marital status and receiving a college degree ($ps > .36$). Availability of informant-reports of women by their husbands/partners was unrelated to receiving a college degree ($ps > .33$). However, coupled (i.e., married or cohabitating) women were more likely to have informant-reports of personality than noncoupled women (76.8% vs. 22.2%, respectively; $\chi^2(1) = 24.45, p < .001$). Availability of self-reports by and informant-reports of men were unrelated to his education ($ps > .18$). However, coupled men were more likely to provide self-reports (77.2% vs. 22.2%, respectively; $\chi^2(1) = 28.14, p < .001$) and have informant-reports (89.8% vs. 72.8%, respectively; $\chi^2(1) = 5.54, p < .05$) than noncoupled men. Availability of informant-reports of men by their wives/ partners was unrelated to marital status or education. No differences in availability of self- or informant-reports for men or women were associated with age, SES, and child race (all $ps > .08$).

As both self- and informant-reports were not available for all women and men, we compared individuals with and without at least one measure of personality to determine if personality data were selectively missing (Table 2). Self-reported Well-Being ratings of women were higher for women with informant-reports than women who did not have informant-reports. No differences were found on the other scales for women or men based on the availability of informant-report data. Similarly, informant-report Well-Being ratings of men who provided self-reports were higher than informant-report ratings of men who did not provide self-reports. No other differences were found on informant-report ratings of men or women based on the availability of self-report data.

Data Analytic Strategy

We took a measurement building approach for our models. Initial models estimated the fit for individual personality dimensions and for self- and informant-reports separately. Subsequently, models were fit sequentially to evaluate measurement invariance. We followed the same logical progression of testing measurement invariance across informants as is used in examinations of longitudinal invariance (Widaman et al., 2010). We estimated models for individual personality dimensions with both self- and informant-reports simultaneously. These models permitted the correlation between the self- and informant-report personality factors to be estimated. Our first test of invariance examined whether the pattern of present (i.e., nonzero) factor loadings is similar across informants (configural invariance). Our second test of invariance examined whether factor loadings for each item are equal across informants (metric invariance) and is tested by constraining the factor loadings across informants to be equal. Our third test of invariance examined whether the probability of item endorsement is similar across informants (scalar invariance; with continuous item responses, these would reflect item means). In the context of the present work that relies on

dichotomous items, we tested scalar invariance by constraining the thresholds across informants to be equal. If configural, metric, and scalar invariance hold, differences in mean trait levels can be interpreted as true score differences, as opposed to being conflated by differences in measurement. A fourth test of invariance is whether the indicator residual variances are equal across informants. However, residual variance equality may have limited clinical and practical utility (Little, Card, Slegers, & Ledford, 2007). Thus, we did not test strict invariance.

All models were estimated using Mplus 7.2 (Muthén & Muthén, 1998–2014). Because observed variables were dichotomous, the weighted least squares estimator with a diagonal weight matrix and robust standard errors and a mean-and variance-adjusted chi-square test statistic (WLSMV) was implemented. This estimation method is the preferred manner of handling dichotomous outcomes in structural equation modeling (Flora & Curran, 2004). Models were evaluated on several indices of goodness of fit, as well as whether or not theoretical predictions, as indicated by specific paths within the model, were supported. To evaluate the overall fit of the models, two indices were selected: the comparative fit index (CFI; Bentler, 1990) and the root mean square error of approximation (RMSEA; Steiger, 1990). Although existing guidelines are somewhat arbitrary (Marsh, Hau, & Wen, 2004), current conventions suggest that excellent fit is indicated by a CFI greater than .95 (Hu & Bentler, 1999) and a RMSEA below or equal to .05 (MacCallum, Browne, & Sugawara, 1996); good fit is indicated by a CFI greater than .90 and a RMSEA between .05 and .10. Where there is conflicting support for evaluating fit between the CFI and RMSEA, we will give preference to RMSEA as this index penalizes less parsimonious models.

In evaluating the presence of measurement invariance, we rely on recommendations from Chen (2007). Based on simulation work with samples greater than 300, Chen (2007) recommends interpreting reductions in CFI of .01 and RMSEA of .015 for testing metric invariance and the same criteria for testing strong invariance. We present chisquare difference tests (as implemented by the DIFFTEST option in Mplus) for completeness, but do not interpret the result as the chi-square difference test is overly sensitive to sample size (Asparouhov & Muthén, 2006).

Results

Initial Models for Personality Dimensions and Reporters Individually

Models were estimated for each personality dimension across each informant individually resulting in a total of 16 models (8 for female and 8 for male targets). Full fit information is provided in Table 3. For all models, all factor loadings to items were significant (unstandardized loadings ranged from .37 to .94, mean loading = .69 [$SD = .12$]; median loading = .70; standardized values ranged from .35 to .95; all $ps < .001$). As our goal was not to assess absolute fit of each dimension, we considered this fit information as a baseline for our model comparison strategy.

Tests of Configural Invariance

Next, we estimated models for each personality scale across informants together. Thus, factors were specified for self-reported and informant-reported personality and the correlation between the self- and informant-report latent personality factors was freely estimated. These analyses yielded a total of 8 models (4 for female and 4 for male targets). Full fit information for these models is provided in Table 4. For all models, all factor loadings to items were significant (unstandardized loadings ranged from .34 to .95, mean loading = .69 [$SD = .12$]; median loading = .71; standardized values ranged from .36 to .94; all $ps < .001$). For all dimensions of personality, associations between latent factors for self- and informant-reports were strong (mean $r = .59$).

Tests of Metric Invariance

The next set of models built on the previous models. In testing metric invariance, the same general two-factor structure was specified. However, factor loadings to the corresponding items were constrained to be equal across the self- and informant-report factors. Rather than testing item pairs individually, equality constraints were placed on the full complement of items to permit a single statistical comparison. Full fit information for these models are provided in Table 4. Across all personality dimensions, there was no evidence that imposing metric invariance constraints resulted in a significant decrement in model fit. For all personality scales, associations between self- and informant-reports were strong (mean $r = .59$). For all models, all factor loadings to items were significant (unstandardized loadings ranged from .36 to .94, mean loading = .70 [$SD = .11$]; median loading = .71; standardized values ranged from .36 to .94; all $ps < .001$).

Tests of Scalar Invariance

Finally, we tested models that added constraints on threshold parameters, in addition to the constraints on the factor loadings. Similar to our statistical comparisons for the factor loadings, we tested comparisons of equality of corresponding item thresholds across self- and informant-reports in a single analysis. Full fit information for these models are provided in Table 4. Overall, most comparisons suggested that introducing constraints on the threshold parameters did not result in a decrement in model fit. No models demonstrated substantial reduction in fit according to the RMSEA and only two of eight models had modest reductions in CFI. As no scales demonstrated a significant decrement in model fit across both fit indices, our data suggest that strong invariance is generally present between self- and informant-reports. Consistent with previously tested models, associations between self- and informant-reports were strong across personality dimensions (mean $r = .59$). For all models, all factor loadings to items were significant (unstandardized loadings ranged from .43 to .94, mean loading = .72 [$SD = .10$]; median loading = .73; standardized values ranged from .36 to .94; all $ps < .001$).

Discussion

The sole reliance on self-reports in psychology is a concern (Podsakoff et al., 2003). Alternative methods of assessment are available and used in some subdisciplines (De Los Reyes & Kazdin, 2005; Durbin et al., 2007; Fox et al., 2005; Hayden et al., 2010; Kessler et

al., 2004; Kessler & Ustün, 2004; Lezak, 2004). In studies of personality, informant-reports have been argued to be a viable alternative assessment method (Vazire, 2006). However, basic psychometric questions about whether the constructs assessed by self- and informant-reports are equivalent have yet to be addressed. The present study examined the psychometric comparability of self- and informant-reports on four personality scales in a sample of adult men and women. We found that the four scales, Well-Being, Social Closeness, Stress Reaction, and Harm Avoidance, demonstrated measurement invariance. Thus, we conclude that these constructs are psychometrically equivalent across self- and informant-reports, and the dimensions they reflect are directly comparable across informants. This suggests that studies can consider using informant-reports as a means of avoiding inflated associations due to relying on a single informant for both personality and the variables personality is hypothesized to predict (Podsakoff et al., 2003). We also failed to find support for the hypothesized differences in relationships between items and constructs across informants due to the observability of the target behaviors in the items and the more unfavorable evaluations provided by informant- than self-report.

Our results supported invariance across structural parameters in the models tested. Nonetheless, similar to previous studies (Connolly et al., 2007; Harkness et al., 1995), we found mean correlations of .59 between self- and informant-reports of personality, indicating that there is a moderate proportion of variance in personality ratings that are unique to each informant. This non-shared variance may have important implications for predicting functional outcomes. Indeed, previous studies report differential associations between self- and informant-reports of personality and a number of psychosocial and mental health outcomes (e.g., Klein, 2003; Oh et al., 2011). However, our results indicate that these differential predictive relationships cannot be attributed to a lack of psychometric comparability between self- and informant-reports. Further work is needed to evaluate how reporter biases and differences in the information available to each source may be related to discrepant reporting between self- and informant-reports.

Although it was not a primary goal of the present study, our work found at least good fit for almost all of the four personality scales examined. The scale that demonstrated the relatively poorest fit was Harm Avoidance for females. Although the MPQ relies on dichotomous responses and primarily includes true–false responses, there are subset of items for the Harm Avoidance scale that uses a different item format. In these items, respondents are asked for preferences between being in a risky situation (e.g., sleeping near rattlesnakes) and an unpleasant, but not risky situation (e.g., walking all day on a blistered foot). Thus, the poorer fit may be due to heterogeneity in response formats. However, this varied response format did not influence self- and informant-reports differently.

The present study benefitted from a reasonably large sample of men and women who provided self- and informant-reports of multiple dimensions of personality. We conducted tests of measurement invariance that are pertinent to determining whether self- and informant-reports of personality are psychometrically comparable. However, our work should be considered in light of a number of limitations. First, we relied on a small number of personality scales from a single inventory. Other traits and instruments may not demonstrate measurement invariance across informants. Second, we focused exclusively on

men and women from married or partnered dyads. Thus, results may differ across different relationships of varying degrees of closeness and duration. Third, as our emphasis was on similarities across informants, we examined men and women as targets of personality assessments separately. As previous studies reported invariant structures of these constructs between men and women, we did not address gender differences here (Marsh et al., 2013; Mor et al., 2008; Reise et al., 2001; Smith & Reise, 1998). However, an advantage of our splitting the sample according to target gender is that it provided a semi-independent replication of our results. Finally, coupled dyads were more likely to provide informant reports than noncoupled dyads. In addition, there were differences between participants with and without data on Well-Being. In our cross-sectional design, it was not possible to determine whether being coupled leads to greater Well-Being or greater Well-Being increases the chances of being coupled. It is conceivable selective attrition associated with lower Well-Being could result in a restriction of range in self-reported Well-Being scores, but, importantly, for both male and female targets, the model showed good fit.

In conclusion, we found that self- and informant-reports of personality reflect the same psychometric constructs. The implications of this work go beyond studies of personality. For example, clinicians and researchers in the area of youth psychopathology typically rely on a combination of self- and informant reports, and are faced with the problem of how to understand the relatively poor agreement between informants (De Los Reyes & Kazdin, 2005; Phares, Compas, & Howell, 1989; Sourander, Helstelä, & Helenius, 1999). As in the personality area, attention to measurement similarities across informants has been neglected. For research in all areas of psychology, it is crucial to examine whether discrepant findings across informants, occasions, subgroups, and cultures are observed for substantive reasons, or whether they can be explained by a lack of psychometric equivalence of the target constructs.

Acknowledgments

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by R01 MH069942 (PI: Klein) and K01 MH092603 (PI: Olino).

References

- Asparouhov, T., Muthén, BO. Robust chi square difference testing with mean and variance adjusted test statistics. Los Angeles: University of California, Los Angeles; 2006.
- Bagby RM, Rector NA, Bindseil K, Dickens SE, Levitan RD, Kennedy SH. Self-report ratings and informants' ratings of personalities of depressed outpatients. *American Journal of Psychiatry*. 1998; 155:437–438. [PubMed: 9501762]
- Bentler PM. Comparative fit indexes in structural models. *Psychological Bulletin*. 1990; 107(2):238–246. [PubMed: 2320703]
- Borkenau P, Riemann R, Angleitner A, Spinath FM. Genetic and environmental influences on observed personality: Evidence from the German Observational Study of Adult Twins. *Journal of Personality and Social Psychology*. 2001; 80:655–668. [PubMed: 11316228]
- Chen FF. Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*. 2007; 14:464–504.

- Church A, Alvarez JM, Mai NT, French BF, Katigbak MS, Ortiz FA. Are cross-cultural comparisons of personality profiles meaningful? Differential item and facet functioning in the Revised NEO Personality Inventory. *Journal of Personality and Social Psychology*. 2011; 101:1068–1089. [PubMed: 21910552]
- 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]
- Connolly JJ, Kavanagh EJ, Viswesvaran C. The convergent validity between self and observer ratings of personality: A meta-analytic review. *International Journal of Selection and Assessment*. 2007; 15:110–117.
- De Los Reyes A, Kazdin AE. Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin*. 2005; 131:483–509. [PubMed: 16060799]
- Durbin CE, Hayden EP, Klein DN, Olino TM. Stability of laboratory-assessed temperamental emotionality traits from ages 3 to 7. *Emotion*. 2007; 7:388–399. [PubMed: 17516816]
- Durbin CE, Wilson S. Convergent validity of and bias in maternal reports of child emotion. *Psychological Assessment*. 2012; 24:647–660. [PubMed: 22149326]
- Flett GL, Besser A, Hewitt PL. Perfectionism, ego defense styles, and depression: A comparison of self-reports versus informant ratings. *Journal of Personality*. 2005; 73:1355–1396. [PubMed: 16138876]
- Flora DB, Curran PJ. An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*. 2004; 9:466–491. [PubMed: 15598100]
- Fox NA, Henderson HA, Marshall PJ, Nichols KE, Ghera MM. Behavioral inhibition: Linking biology and behavior within a developmental framework. *Annual Review of Psychology*. 2005; 56:235–262.
- Funder DC. Accurate personality judgment. *Current Directions in Psychological Science*. 2012; 21:177–182.
- Harkness AR, Tellegen A, Waller N. Differential convergence of self-report and informant data for Multidimensional Personality Questionnaire traits: Implications for the construct of negative emotionality. *Journal of Personality Assessment*. 1995; 64:185–204. [PubMed: 16367738]
- Hayden EP, Durbin CE, Klein DN, Olino TM. Maternal personality influences the relationship between maternal reports and laboratory measures of child temperament. *Journal of Personality Assessment*. 2010; 92:586–593. [PubMed: 20954060]
- Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*. 1999; 6:1–55.
- John OP, Robins RW. Accuracy and bias in self-perception: individual differences in self-enhancement and the role of narcissism. *Journal of Personality and Social Psychology*. 1994; 66:206–219. [PubMed: 8126650]
- Johnson W, Spinath F, Krueger RF, Angleitner A, Riemann R. Personality in Germany and Minnesota: An IRT-based comparison of MPQ self-reports. *Journal of Personality*. 2008; 76:665–706. [PubMed: 18399949]
- Kessler RC, Abelson J, Demler O, Escobar JJ, Gibbon M, Guyer ME, Walters EE. Clinical calibration of DSM-IV diagnoses in the World Mental Health (WMH) version of the World Health Organization (WHO) Composite International Diagnostic Interview (WMH-CIDI). *International Journal of Methods in Psychiatric Research*. 2004; 13:122–139. [PubMed: 15297907]
- Kessler RC, Üstün TB. The World Mental Health (WMH) Survey Initiative Version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). *International Journal of Methods in Psychiatric Research*. 2004; 13:93–121. [PubMed: 15297906]
- Klein DN. Patients' versus informants' reports of personality disorders in predicting 7 1/2-year outcome in outpatients with depressive disorders. *Psychological Assessment*. 2003; 15:216–222. [PubMed: 12847782]
- Laverdiere O, Morin AJ, St-Hilaire F. Factor structure and measurement invariance of a short measure of the Big Five personality traits. *Personality and Individual Differences*. 2013; 55:739–743.

- Lezak, MD. Neuropsychological assessment. New York, NY: Oxford University Press; 2004.
- Little, TD., Card, NA., Slegers, DW., Ledford, EC. Representing contextual effects in multiple-group MACS models. In: Little, TD., Bovaird, JA., Card, NA., editors. Modeling contextual effects in longitudinal. Mahwah, NJ: Erlbaum; 2007. p. 121-147.
- MacCallum RC, Browne MW, Sugawara HM. Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*. 1996; 1:130-149.
- Markon KE, Quilty LC, Bagby RM, Krueger RF. The development and psychometric properties of an informant-report form of the Personality Inventory for DSM-5 (PID-5). *Assessment*. 2013; 20:370-383. [PubMed: 23612961]
- Marsh HW, Hau KT, Wen Z. In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*. 2004; 11:320-341.
- Marsh HW, Ludtke O, Muthén B, Asparouhov T, Morin AJ, Trautwein U, Nagengast B. A new look at the big five factor structure through exploratory structural equation modeling. *Psychological Assessment*. 2010; 22:471-491. [PubMed: 20822261]
- Marsh HW, Nagengast B, Morin AJ. Measurement invariance of big-five factors over the life span: ESEM tests of gender, age, plasticity, maturity, and la dolce vita effects. *Developmental Psychology*. 2013; 49:1194-1218. [PubMed: 22250996]
- McCrae RR, Costa PT, Hrebíková M, Urbánek T, Martin TA, Oryol VE, Senin IG. Age differences in personality traits across cultures: Self-report and observer perspectives. *European Journal of Personality*. 2004; 18:143-157.
- Merz EL, Malcarné VL, Roesch SC, Ko CM, Emerson M, Roma VG, Sadler GR. Psychometric properties of Positive and Negative Affect Schedule (PANAS) original and short forms in an African American community sample. *Journal of Affective Disorders*. 2013; 151:942-949. [PubMed: 24051099]
- 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]
- Mor N, Zinbarg RE, Craske MG, Mineka S, Uliaszek A, Rose R, Waters AM. Evaluating the invariance of the Factor Structure of the EPQ-R-N among adolescents. *Journal of Personality Assessment*. 2008; 90:66-75. [PubMed: 18444097]
- Muthén, LK., Muthén, BO. *Mplus user's guide*. 7th. Los Angeles, CA: Muthén & Muthén; 1998-2014.
- Oh I-S, Wang G, Mount MK. Validity of observer ratings of the five-factor model of personality traits: A meta-analysis. *Journal of Applied Psychology*. 2011; 96:762-773. [PubMed: 21142341]
- Olino TM, Klein DN, Dyson MW, Rose SA, Durbin CE. Temperamental emotionality in preschool-aged children and depressive disorders in parents: Associations in a large community sample. *Journal of Abnormal Psychology*. 2010; 119:468-478. [PubMed: 20677836]
- Patrick CJ, Curtin JJ, Tellegen A. Development and validation of a brief form of the Multidimensional Personality Questionnaire. *Psychological Assessment*. 2002; 14:150-163. [PubMed: 12056077]
- Phares V, Compas BE, Howell DC. Perspectives on child behavior problems: Comparisons of children's self-reports with parent and teacher reports. *Psychological Assessment*. 1989; 1:68-71.
- Podsakoff PM, MacKenzie SB, Lee J-Y, Podsakoff NP. Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*. 2003; 88:879-903. [PubMed: 14516251]
- Ready RE, Clark LA. Correspondence of psychiatric patient and informant ratings of personality traits, temperament, and interpersonal problems. *Psychological Assessment*. 2002; 14:39-49. [PubMed: 11911048]
- Reise SP, Smith L, Furr RM. Invariance on the NEO PI-R neuroticism scale. *Multivariate Behavioral Research*. 2001; 36:83-110.
- Riso LP, Klein DN, Anderson RL, Ouimette PC, Lizardi H. Concordance between patients and informants on the personality disorder examination. *American Journal of Psychiatry*. 1994; 151:568-573. [PubMed: 8147455]

- Rowinski T, Ciecuch J, Oakland T. The factorial structure of four temperament styles and measurement invariance across gender and age groups. *Journal of Psychoeducational Assessment*. 2014; 32:77–82.
- Smith LL, Reise SP. Gender differences on negative affectivity: An IRT study of differential item functioning on the Multidimensional Personality Questionnaire Stress Reaction scale. *Journal of Personality and Social Psychology*. 1998; 75:1350–1362. [PubMed: 9866192]
- Sourander A, Helstelä L, Helenius H. Parent-adolescent agreement on emotional and behavioral problems. *Social Psychiatry and Psychiatric Epidemiology*. 1999; 34:657–663. [PubMed: 10703276]
- South SC, Oltmanns TF, Johnson J, Turkheimer E. Level of agreement between self and spouse in the assessment of personality pathology. *Assessment*. 2011; 18:217–226. [PubMed: 21220382]
- Spence R, Owens M, Goodyer I. The longitudinal psychometric properties of the EAS Temperament Survey in adolescence. *Journal of Personality Assessment*. 2013; 95:633–639. [PubMed: 23905630]
- Steiger JH. Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*. 1990; 25:173–180. [PubMed: 26794479]
- Vandenberg RJ, Lance CE. A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*. 2000; 3:4–70.
- Vazire S. Informant reports: A cheap, fast, and easy method for personality assessment. *Journal of Research in Personality*. 2006; 40:472–481.
- Vazire S. Who knows what about a person? The self- other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology*. 2010; 98:281–300. [PubMed: 20085401]
- Widaman KF, Ferrer E, Conger RD. Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives*. 2010; 4(1): 10–18. [PubMed: 20369028]
- Woo SE, Chernyshenko OS, Longley A, Zhang Z-X, Chiu C-Y, Stark SE. Openness to experience: Its lower level structure, measurement, and cross-cultural equivalence. *Journal of Personality Assessment*. 2014; 96:29–45. [PubMed: 23795950]

Table 1

Self- and Informant-Report Personality Descriptive Statistics.

	Self-report			Informant-report		
	α	Mean	SD	α	Mean	SD
Females						
Well-Being	.81	10.25	3.03	.87	9.75	3.73
Social Closeness	.77	9.40	2.49	.84	9.60	2.85
Stress Reaction	.85	6.41	4.02	.86	6.00	4.14
Harm Avoidance	.67	9.72	2.12	.69	9.66	2.23
Males						
Well-Being	.82	9.99	3.29	.89	9.44	4.03
Social Closeness	.86	7.81	3.41	.87	7.44	3.57
Stress Reaction	.86	4.57	3.90	.87	4.71	3.95
Harm Avoidance	.76	7.71	2.88	.82	7.61	3.24

Table 2

Personality Comparisons Between Women and Men With and Without Informant-Reports and Self-Reports.

	Without informant-report		With informant-report		<i>t</i>	Without self-report		With self-report		<i>t</i>
	<i>M</i> (<i>SD</i>)	<i>n</i>	<i>M</i> (<i>SD</i>)	<i>n</i>		<i>M</i> (<i>SD</i>)	<i>n</i>	<i>M</i> (<i>SD</i>)	<i>n</i>	
Females										
Well-Being	9.55 (3.13)	<i>n</i> = 86	10.40 (3.00)	<i>n</i> = 386	-2.38*	7.92 (4.01)	<i>n</i> = 9	9.80 (3.71)	<i>n</i> = 385	-1.49
Social Closeness	9.22 (2.52)		9.44 (2.49)		-0.74	8.22 (4.63)		9.63 (2.8)		-0.91
Stress Reaction	6.31 (4.44)		6.43 (3.93)		-0.23	7.00 (5.61)		5.98 (4.1)		0.73
Harm Avoidance	9.81 (1.97)		9.70 (2.16)		0.42	8.97 (3.07)		9.67 (2.2)		-0.69
Males										
Well-Being	10.40 (3.41)	<i>n</i> = 10	9.98 (3.29)	<i>n</i> = 387	0.40	8.37 (4.14)	<i>n</i> = 86	9.66 (3.97)	<i>n</i> = 386	-2.61**
Social Closeness	7.20 (4.26)		7.82 (3.39)		-0.57	7.33 (3.74)		7.46 (3.54)		-0.29
Stress Reaction	3.11 (2.96)		4.61 (3.92)		-1.19	5.18 (3.82)		4.61 (3.97)		1.17
Harm Avoidance	6.87 (2.58)		7.73 (2.89)		-0.93	7.75 (2.97)		7.58 (3.30)		0.43

* $p < .05$.

** $p < .01$.

Table 3

Summary of Model Fit for Self- and Informant-Reports Personality.

	Self-report				Informant-report			
	χ^2	df	CFI	RMSEA	χ^2	df	CFI	RMSEA
Females								
Well-Being	155.61	77	.96	.05 (.04–.06)	169.91	77	.98	.05 (.04–.07)
Social Closeness	117.51	54	.95	.05 (.04–.06)	87.69	54	.98	.04 (.02–.05)
Stress Reaction	434.62	90	.90	.09 (.08–.10)	424.95	90	.91	.10 (.09–.11)
Harm Avoidance	107.01	54	.90	.05 (.03–.06)	131.75	54	.87	.06 (.05–.07)
Males								
Well-Being	155.74	77	.96	.05 (.04–.06)	258.91	77	.96	.07 (.06–.08)
Social Closeness	132.54	54	.97	.06 (.05–.07)	163.59	54	.98	.07 (.05–.08)
Stress Reaction	288.70	90	.95	.07 (.06–.08)	418.99	90	.93	.09 (.08–.10)
Harm Avoidance	119.66	54	.94	.05 (.04–.07)	174.04	54	.94	.07 (.06–.08)

Note. *df* = degrees of freedom; CFI = comparative fit index; RMSEA = root mean square error of approximation. RMSEA values are presented with point estimates and 90% confidence intervals in parentheses.

Table 4

Model Fit Statistics for Evaluating Measurement Invariance.

	χ^2	df	CFI	RMSEA	$r_{\text{self-informant}}$	Model test	χ^2
Females							
Well-Being							
I. Configural Invariance	465.27	349	0.978	.026 (.020–.032)	.52***	—	
II. Metric Invariance	513.1	363	0.972	.029 (.023–.035)	.51***	II vs. I	38.21***
III. Metric & Scalar Invariance	606.92	377	0.957	.036 (.030–.041)	.51***	III vs. II	142.20***
Social Closeness							
I. Configural Invariance	360.54	251	0.966	.030 (.023–.037)	.65***	—	
II. Metric Invariance	399.58	263	0.958	.033 (.026–.039)	.64***	II vs. I	32.87**
III. Metric & Scalar Invariance	443.65	275	0.948	.036 (.029–.042)	.64***	III vs. II	64.49***
Stress Reaction							
I. Configural Invariance	965.24	404	0.906	.054 (.049–.058)	.46***	—	
II. Metric Invariance	908.24	419	0.918	.049 (.045–.054)	.46***	II vs. I	18.04
III. Metric & Scalar Invariance	962.74	434	0.911	.050 (.046–.055)	.46***	III vs. II	74.11***
Harm Avoidance							
I. Configural Invariance	411.36	251	0.853	.036 (.030–.043)	.63***	—	
II. Metric Invariance	405.5	263	0.869	.034 (.027–.040)	.62***	II vs. I	13.69
III. Metric & Scalar Invariance	426.63	275	0.861	.034 (.027–.040)	.62***	III vs. II	26.76**
Males							
Well-Being							
I. Configural Invariance	581.9	349	0.962	.037 (.032–.043)	.52***	—	
II. Metric Invariance	624.37	363	0.958	.039 (.034–.044)	.49***	II vs. I	43.58***
III. Metric & Scalar Invariance	655.61	377	0.953	.040 (.035–.045)	.49***	III vs. II	56.23***
Social Closeness							
I. Configural Invariance	477.63	251	0.970	.044 (.038–.049)	.73***	—	
II. Metric Invariance	442.56	263	0.976	.038 (.032–.044)	.73***	II vs. I	14.25

	χ^2	df	CFI	RMSEA	$r_{\text{self-informant}}$	Model test	χ^2
III. Metric & Scalar Invariance	478.27	275	0.973	.039 (.033–.045)	.73 ***	III vs. II	48.75 ***
Stress Reaction							
I. Configural Invariance	908.75	404	0.93	.051 (.047–.056)	.54 ***	—	
II. Metric Invariance	842.27	419	0.941	.046 (.042–.051)	.54 ***	II vs. I	17.65
III. Metric & Scalar Invariance	901.33	434	0.935	.048 (.043–.052)	.53 ***	III vs. II	85.97 ***
Harm Avoidance							
I. Configural invariance	505.81	251	0.926	.046 (.040–.052)	.73 ***	—	
II. Metric Invariance	510.76	263	0.928	.044 (.039–.050)	.71 ***	II vs. I	29.22 **
III. Metric & Scalar Invariance	568.79	275	0.909	.049 (.043–.054)	.71 ***	III vs. II	110.53 ***

Note. df = degrees of freedom; CFI = comparative fit index; RMSEA = root mean square error of approximation. RMSEA values are presented with point estimates and 90% confidence intervals in parentheses. All models included correlated factors for self- and informant reports. Tests of configural invariance estimate models without constraints on factor loadings or threshold parameters; tests of metric invariance estimate models with equality constraints on factor loadings, but not on threshold parameters, across informants; and tests of scalar invariance estimate models with equality on factor loadings and threshold items across informants.

 $p < .01$.

 $p < .001$.