



Calibrating personality self-report scores to acquaintance ratings

John A. Johnson¹

Department of Psychology, Pennsylvania State University, University Park, PA 16803, United States

ARTICLE INFO

Keywords:

Cutting scores
Interpreting personality scores
Optimal Data Analysis

ABSTRACT

By convention in individual personality assessment, scores on self-report questionnaires within $\pm .5$ standard deviations of the mean score for that trait are considered "average," whereas scores outside that range are reported as "high" or "low" levels of the trait. To date, no one has examined how well this convention corresponds to perceptions of low, average, or high trait levels by acquaintances. The present study compares the accuracy of the conventional $\pm .5$ SD cutting scores for low, average, and high trait levels with cutting scores derived from Optimal Data Analysis (ODA; Yarnold & Seltysik, 2005) for 160 participants who completed the IPIP-NEO-300 (Goldberg, 1999; Johnson, 2014) and were rated by acquaintances on the 30 facets and 5 domains of that instrument. In 32 of 35 comparisons, the $\pm .5$ convention was found to be less accurate than the ODA method for deriving cutting scores.

1. Calibrating personality self-report scores to acquaintance ratings

A perennial practical problem when providing feedback on personality inventories has been how to choose definitive descriptions of personality from scores that are merely probabilistic signs of personality. Hofstee and de Raad (1992) have argued that, given the imperfect reliabilities of personality measures, we should not pretend that our feedback to inventory respondents can be more precise than to tell them that they possess "low," "average," or "high" levels of a personality trait. Often this is accomplished by labeling scores within $\pm .5$ standard deviations of the mean for a standardization sample as *average* and scores outside that range as *low* or *high*. This means that about 40% of the population would be described as "average" for normally distributed traits. The research described in this article tested the validity of this practice. Specifically, the research examined the correspondence between "low," "average," and "high" levels of traits, defined by the $\pm .5$ SD convention for self-reports and ratings given by acquaintances based explicitly on the categories *low*, *average*, and *high*.

Examples of widely used, multi-scale personality inventories that have employed the $\pm .5$ SD definition of *average* include the California Psychology Inventory (CPI; Gough & Bradley, 1996), Hogan Personality Inventory (HPI, Hogan & Hogan, 1992), NEO Personality Inventory (NEO-PI; Costa & McCrae, 1992), and the 16PF (IPAT Staff, 1979). All of these inventories have used profile interpretation sheets that graph scores along standardized scales such as *T* scores, percentiles, or

standardized tens (stems) to provide a visual representation of which scores are inside or outside of the average range. Despite some variation across the inventories and even within different interpretive guides for each inventory, there is considerable consensus on the $\pm .5$ standard deviation standard for defining *average*.

Profile sheets for the CPI (Gough & Bradley, 1987) graph results in terms of standardized *T* scores. An interpretive guide for the CPI (McAllister, 1986) classifies *T* scores below 30 as *very low*, 30–45 as *low*, 55–70 as *high*, and above 70 as *very high*. This implicates scores between 45 and 55 (i.e., within $\pm .5$ SD of the mean) as *average*. Hogan and Hogan (1992, p. 39) are quite explicit about the $\pm .5$ SD designation of *average*, "Our rule of thumb is to call scores above the 70th percentile high and scores below the 30th percentile low." (The 70th and 30th percentiles correspond to $\pm .5$ SDs for normally distributed scores; see Fig. 1).

The one-page "Your NEO Summary" feedback sheet (Costa & McCrae, 1992, p. 9), uses three descriptions for each NEO domain scale, where "*T* scores of 56 or higher are considered high, *T* scores ranging from 55 to 45 are considered average, and *T* scores of 44 or lower are considered low" (p. 7). However, for the professional interpretive profile sheet, Costa and McCrae (1992, p. 20), add two additional categories, *very low* (*T* scores below 35) and *very high* (*T* scores above 65). For the 16PF, "Stems 5 and 6 extend, respectively, a half standard deviation below and above the mean, constituting the solid center of the population" (IPAT Staff, 1979, p. 17). This area is shaded in gray on the 16PF test profile sheet (IPAT Staff, 1973), although the IPAT Staff

E-mail address: j5j@psu.edu.

¹ John A. Johnson, Pennsylvania State University. Some of the findings in this article were presented in at the first stand-alone conference of the Association for Research in Personality, Evanston, IL, July 2009.

<https://doi.org/10.1016/j.paid.2019.109734>

Received 7 October 2019; Accepted 26 November 2019

Available online 04 December 2019

0191-8869/© 2019 Elsevier Ltd. All rights reserved.

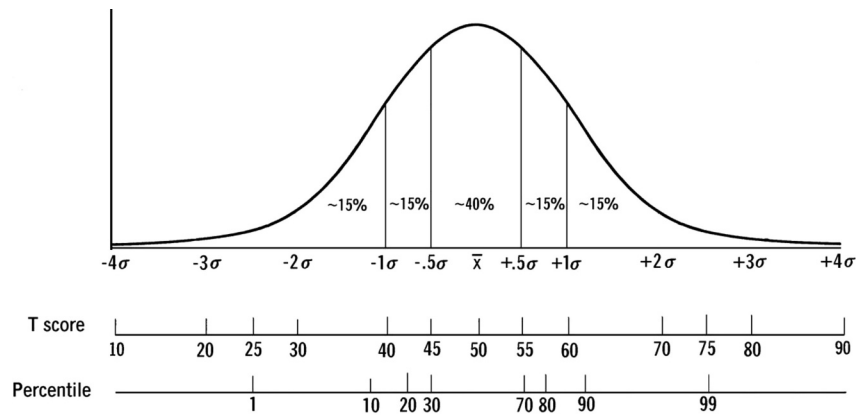


Fig.. 1. Percent of population within .5 SD of the Mean, 1 SD, and beyond for normally distributed scores.

(1979, p. 17) also suggests that "One would normally consider sten scores of 5 through 7 [i.e., scores within ± 1 SD] to be average."

The interpretive guidelines from these four personality inventories, although agreeing that scores within $\pm .5$ SD of the mean are *average*, also reflect the arbitrariness of interpretive cutting scores. The IPAT Staff calls the $\pm .5$ SD area of the score distribution the "solid center" and highlights that range on the 16PF profile sheet, but also indicates that perhaps scores within ± 1 SD of the mean (67% rather than ~40% of normally distributed scores) can be considered average. There are also differences of opinion on whether scores outside the average range might be subdivided into *low* versus *very low* and *high* versus *very high* and where the cutting scores for those further subdivisions should be. It would seem that some kind of objective criterion might be helpful in sorting out what amounts to arbitrary cutting scores (Blanton & Jaccard, 2006).

Hofstee and de Raad (1992) suggest calculating how much scores would have to differ to be to be considered significantly different in the traditional statistical sense. They note that, given the typical reliability for personality scales, a score must be $\pm .14$ standard deviations from the mean to be significantly different ($p < .05$) from the mean. However, Hofstee and de Raad's cutting scores would place about 86% of all persons in the *average* category with only 7% described as *low* and 7% as *high*. One is left wondering if only 14% of the population is distinguishably different enough to be labeled with a particular personality trait. Some other method of objectivity seems to be needed.

Wim Hofstee has suggested two additional objective methods. One is to take anchor points on personality item response scales literally rather than following the tradition of scaling them according to the mean for a standardization sample (Hofstee, 2009; Hofstee & Hendricks, 1998). Because this idea has not gained wide acceptance and was not used in the current study, it will be only briefly described. The rationale for accepting at face value the anchor labels of a five-point response scale can be demonstrated with an example. If a person's response to the item "I like cod-liver oil" is "disagree" (2 on a 5-point scale of agreement) but the mean response for the item is 1.5, "It is as if you are being told that you actually like cod-liver oil because the average person despises it even more than you do" (Hofstee, 2009, p. 229). Hofstee therefore recommends that for the purpose of describing an individual, item responses be interpreted in terms of deviations from the midpoint rather than a sample's mean response to the item.

Another method offered by Hofstee for objective personality assessment has been much more widely accepted: deferring to averaged assessments from persons who are well-acquainted with a person (Hofstee, 1994, 2009). This argument is based on a fundamental idea in measurement, that averaged multiple observations are inevitably more reliable than a single measurement (Block, 1978). Whereas self-reports necessarily rely on the judgments of one person, acquaintance ratings can be averaged into composites that are more reliable than self-

reports. Furthermore, notes Hofstee (2009), acquaintances have a perspective that allows the assessment of traits that are difficult or impossible to convey by self-report: "Take [the third-person item] 'he/she overestimates himself/herself'. In the first person, the answer would be paradoxical: If I agree that I overestimate myself, I am no longer overestimating myself; if I disagree, one might take that as just another proof of the proposition" (pp. 227–228). A similar paradox occurs when someone agrees with self-report item such as "I am more humble than most people." Even in cases where no paradox is involved, respondents can simply find it difficult to assess themselves objectively. Although complete objectivity may be impossible in either self-report or acquaintance judgment, the closest route to objectivity is the intersubjectivity of averaged judgments from multiple judges, wherein idiosyncrasies are hopefully canceled out (Block, 1978; Hofstee, 1994, 2009; Kolar, Funder & Colvin, 1996).

The potentially greater objectivity of acquaintance judgments over self-judgment suggests examining how far scores on self-report personality scales must deviate from the mean to be judged by acquaintances as below or above average. It is a simple empirical question how accurately the $\pm .5$ SD cutting scores or any other method for choosing cutting scores for a particular self-report measure will correspond to judgments of personality trait elevation by acquaintances. To the author's knowledge no such empirical data have been published, although research comparing self- and acquaintance ratings by Q-sort has indicated that people tend to give themselves higher ratings for favorable traits and less-observable, inner traits than acquaintances (Funder & Colvin, 1997). Such mean differences could translate into differences between self- and other judgments of whether someone shows a low, average, or high level of a trait.

The first goal of the current research was to simply observe and describe the accuracy of the $\pm .5$ SD convention for predicting acquaintance judgments by noting how frequently persons classified as low, average, and high by self-report were classified the same way by acquaintance judgment. The second goal was to compare accuracy of the $\pm .5$ SD against an alternative method for deriving cutting scores for maximizing valid classification called *Optimal Data Analysis* (ODA; Yarnold & Soltysik, 2005). ODA is simply an improvement of long-standing but suboptimal heuristics for classifying individuals with test scores (e.g. Dawes & Meehl, 1966; Rorer, Hoffman & Hsieh, 1966) that leverages modern computing power for iteration that was unavailable in classic methods. Further details on ODA are provided in the Methods section.

2. Method

196 undergraduate students agreed to complete the IPIP-NEO-300 (Goldberg, 1999), a 300-item representation of Costa and McCrae's (1992) NEO PI-R. Each facet scale of the IPIP-NEO-300

contain 10 items in contrast to the NEO PI-R facet scales, which contain eight items. Consequently, alpha reliabilities for the IPIP-NEO-300 slightly exceed alpha reliabilities for the NEO PI-R (Goldberg, 1999). The IPIP-NEO-300 scales correlate .86 on average with their corresponding NEO PI-R scales, corrected for attenuation (Goldberg, 1999). Continued research has strongly supported the reliability and validity of the IPIP-NEO-300 scales (Johnson, 2014). The Jackson split-half individual reliability and Goldberg psychometric antonym indices of protocol validity (Johnson, 2005) verified that participants were responding attentively and appropriately.

Acquaintance judgments of personality were gathered following the methods described by Vazire (2006). Participants who completed the IPIP-NEO-300 provided the names of three knowledgeable acquaintances, who were invited to a Web site containing descriptions of the five domains and 30 facets measured by the IPIP-NEO. Acquaintances were asked to rate participants on 35 scales with the following percentile anchor points: 1, 10, 20, 30, 40, 50, 60, 70, 80, 90, 99. From the 196 persons recruited, 160 (59 males, 101 females) actually completed the IPIP-NEO and had at least one acquaintance complete the personality ratings (116 had three raters, 39 had two raters, and 5 had one rater). For illustrative purposes, Fig. 2 shows a portion of the Web rating form.

Acquaintance ratings were averaged for participants with two or more raters to increase reliability. Intraclass correlation estimates of reliability, presented in Table 1, ranged from .14 to .58, with an average ICC of .44. As a check on the overall validity of the self-reports and acquaintance ratings, averaged acquaintance percentile ratings were correlated with the self-report IPIP-NEO scores. These results, also reported in Table 2, show that the self-acquaintance correlations ranged from $r = .14$ to $.61$, with an average r of $.37$, which is typical for correlations between self and friends or family members (Connelly & Ones, 2010). These results indicate that the participants took the task seriously enough to proceed to the main analyses, the prediction of low, average, and high acquaintance ratings from self-report.

In accordance with the research plan for the current study, averaged acquaintance ratings in the 30–70 percentile range were classified as "average" and ratings outside that range as "low" or "high." (More fine-grained analyses of the "very low" and "very high" categories were left to future research). The three categories defined by acquaintance ratings became the target of prediction from IPIP-NEO-300 self-report scores for the 30 facets and five domains.

Percent of correct classifications for the 35 ratings were compared for the two methods of deriving cutting scores ($\pm .5$ SD convention versus ODA). IPIP-NEO-300 means and SDs from Johnson's (2005) standardization sample of over 20,000 persons were used to describe whether IPIP-NEO-300 scale scores were with $.5$ SD of the mean. ODA

uses an iterative process to determine cutting scores that maximize correct classifications, employing a leave-one-out procedure (LOO) to confirm the stability of the predictive model. LOO is conceptually equivalent to cross-validation of the derived cutting scores. (See Yarnold & Soltysik, 2005, for further explanation.)

For seven traits, insufficient cases in a category precluded the LOO procedure. Although ODA produces many statistics, including overall percentage accuracy in classification (PAC), sensitivity (PAC within each category), predictive value (percent of persons correctly classified), effect strength for sensitivity and predictive value (classification improvement over chance), and others (including model efficiency over base rates), for current analyses were limited to overall PAC, which was compared to PAC for the $\pm .5$ SD method.

3. Results

Examination of averaged acquaintance rating frequencies showed that, as logic might suggest, more persons were rated as *average* than as *low* or *high* for all personality traits. The frequency of persons rated in the average category ranged from 48% (Depression) to 82% (Liberalism). However, frequencies outside of the average category were highly skewed away from what is generally considered to be the less socially desirable end of each trait dimension (introversion, disagreeableness, unconscientiousness, neuroticism, closedness to experience). Only between 1% (for low Agreeableness and Conscientiousness and several facets of those two traits) and 11% (for both low Openness to Experience and Liberalism) of the sample was rated by acquaintances in the less desirable category. In contrast, the frequency of persons predicted to fall in the less desirable category by the $\pm .5$ SD convention ranged from 17% for low Gregariousness to 63% for overall low Openness to Experience. Thus, the $\pm .5$ SD method appeared to overestimate how many persons would be judged by acquaintances to fall into a less socially desirable personality category. The ODA method, which is calibrated on the distribution of acquaintance ratings, should be less likely to make such miscategorizations.

The superior accuracy of ODA over the traditional $\pm .5$ SD cutting score method is borne out in Table 1. For 32 of 35 traits, the ODA method of selecting cutoffs produced a greater overall percentage accuracy of classification (PAC). The mean PAC for the ODA method was 62.2% ($SD = 12.9$), whereas the mean PAC for the $\pm .5$ SD method was 41% ($SD = 3.8$); $t(34) = 21.2$, $p < .0001$. For the ODA method, optimal low cutting scores ranged from $.17$ SD above the mean (for overall Agreeableness) to 2.4 SD below the mean (for overall Extraversion), with an average of 1.21 SD below the mean. High cutting scores ranged from $.03$ SD below the mean (for Dutifulness) to 2.31 SD above the mean (for Vulnerability), with an average of 1.27 SD above the mean.

Extraversion

Extraversion is marked by pronounced engagement with the external world. Extraverts enjoy being with people, are full of energy, and often experience positive emotions. They tend to be enthusiastic, action-oriented, individuals who are likely to say "Yes!" or "Let's go!" to opportunities for excitement. In groups they like to talk, assert themselves, and draw attention to themselves.

Introverts lack the exuberance, energy, and activity levels of extraverts. They tend to be quiet, low-key, deliberate, and disengaged from the social world. Their lack of social involvement should not be interpreted as shyness or depression; the introvert simply needs less stimulation than an extravert and prefers to be alone. The independence and reserve of the introvert is sometimes mistaken as unfriendliness or arrogance. In reality, an introvert who is high on the agreeableness dimension will not seek others out but will be quite pleasant when approached.

Friendliness: Friendly people genuinely like other people and openly demonstrate positive feelings toward others. They make friends quickly and it is easy for them to form close, intimate relationships. Low levels of Friendliness do not necessarily indicate coldness and hostility, but people with low Friendliness do not reach out to others and are perceived as distant and reserved.

1	10	20	30	40	50	60	70	80	90	99
Very Low		Low-Average		Middle-Average		High-Average		Very High		

Gregariousness: Gregarious people find the company of others pleasantly stimulating and rewarding. They enjoy the excitement of crowds. Low-gregarious individuals tend to feel overwhelmed by, and therefore actively avoid, large crowds. They do not necessarily dislike being with people sometimes, but their need for privacy and time to themselves is much greater than for individuals who are high on gregariousness.

1	10	20	30	40	50	60	70	80	90	99
Very Low		Low-Average		Middle-Average		High-Average		Very High		

Assertiveness: Individuals high on Assertiveness like to speak out, take charge, and direct the activities of others. They tend to be leaders in groups. Persons low on assertiveness tend not to talk much and let others control the activities of groups.

1	10	20	30	40	50	60	70	80	90	99
Very Low		Low-Average		Middle-Average		High-Average		Very High		

Fig. 2. Portion of the Web rating form for assessing three facets of extraversion.

Table 1
Acquaintance rating reliabilities and correlations with IPIP-NEO-300 Self-Reports.

IPIP-NEO-300 Self-Report																		
ACQ	IOC	E1	E2	E3	E4	E5	E6	E	A1	A2	A3	A4	A5	A6	A7	C1	C2	C3
E1	.49	.41	.34	.26	.13	.12	.35	.38	.15	-.03	.11	.06	-.16	.04	.04	.12	.11	.10
E2	.38	.39	.42	.35	.12	.27	.31	.45	.09	-.15	.00	-.12	-.27	-.08	-.12	.10	.03	-.04
E3	.58	.22	.27	.47	.32	.17	.08	.36	-.03	-.09	-.04	-.22	-.24	-.04	-.15	.21	.04	-.04
E4	.52	.20	.24	.37	.03	.12	.26	.11	.01	.01	.04	.01	-.19	.06	.01	.22	.08	.00
E5	.54	.23	.33	.35	.21	.46	.17	.43	.00	-.35	-.21	-.30	-.33	-.26	-.32	.09	-.11	-.27
E6	.50	.31	.22	.17	.11	.04	.39	.29	.25	.05	.17	.17	-.11	.11	.14	.20	.05	.10
E	.51	.36	.37	.36	.21	.23	.28	.43	.09	-.12	.04	-.09	-.25	-.04	-.08	.11	.04	-.06
A1	.42	.24	.14	.03	.02	.02	.33	.18	.30	.02	.13	.10	.00	.13	.15	.10	.12	.18
A2	.14	.10	.01	.06	.15	-.15	.20	.04	.11	.14	.08	.19	.02	.16	.16	.16	.05	.20
A3	.24	.10	-.05	-.08	.10	-.17	.07	-.02	.14	.19	.16	.22	.11	.25	.24	.06	.14	.25
A4	.19	-.03	-.16	-.16	-.06	-.23	.06	-.15	.15	.16	.16	.25	.05	.19	.22	.05	.22	.22
A5	.24	.05	-.07	-.15	-.11	-.16	.06	-.09	.15	.23	.12	.23	.22	.20	.26	.02	.05	.22
A6	.39	.21	.08	-.16	-.03	-.20	.21	.02	.27	.32	.34	.28	.18	.44	.40	.02	.18	.30
A	.41	.12	.00	-.09	-.04	-.15	.23	.01	.24	.18	.24	.27	.10	.28	.29	.11	.13	.23
C1	.50	.08	-.02	.18	.22	-.01	.05	.10	.04	.02	-.11	.04	-.15	-.04	-.04	.26	.18	.09
C2	.53	.03	-.11	-.06	.23	-.13	.03	-.03	.13	.15	.08	.19	.11	.18	.19	.17	.57	.22
C3	.39	.06	-.15	-.06	.20	-.19	.05	-.05	.13	.25	.17	.16	.14	.24	.24	.16	.28	.35
C4	.49	.02	-.14	.03	.18	-.09	.02	-.02	.08	.08	-.01	.09	-.02	.06	.07	.24	.31	.22
C5	.54	.05	-.04	.03	.24	-.05	.02	.04	.03	.15	.00	.17	.05	.10	.12	.25	.33	.23
C6	.43	-.02	-.16	-.12	.05	-.21	.02	-.13	.08	.25	.12	.24	.21	.21	.25	.16	.24	.32
C	.52	-.03	-.18	-.03	.13	-.15	.01	-.08	.06	.19	.03	.18	.06	.15	.15	.25	.36	.31
N1	.36	-.25	-.21	-.28	-.10	-.30	-.25	-.34	-.10	.12	.04	.13	.23	.18	.14	-.26	.03	-.04
N2	.54	-.17	-.07	.09	-.09	.05	-.24	-.09	-.26	-.10	-.17	-.31	-.02	-.19	-.24	-.19	-.02	-.21
N3	.51	-.31	-.22	-.24	-.15	-.16	-.36	-.34	-.21	.07	-.01	-.02	.22	.18	.05	-.36	.06	-.14
N4	.52	-.31	-.29	-.38	-.21	-.24	-.26	-.40	-.10	.12	.05	.12	.26	.24	.16	-.31	-.02	-.06
N5	.39	-.05	.03	.02	-.15	.11	-.19	-.03	-.16	-.15	-.04	-.18	-.03	-.03	-.14	-.35	-.14	-.32
N6	.51	-.14	-.03	-.27	-.18	-.08	-.16	-.19	-.07	.13	.11	.10	.24	.21	.16	-.35	-.01	-.06
N	.46	-.27	-.20	-.26	-.23	-.14	-.32	-.33	-.11	.07	.03	-.03	.22	.11	.06	-.35	-.03	-.15
O1	.47	.01	.04	.10	-.18	.04	.10	.04	.00	-.15	-.06	-.04	-.09	.06	-.06	-.06	-.08	-.15
O2	.52	.05	.03	.05	.02	-.02	.09	.05	.02	.03	.03	.11	-.01	.20	.09	.04	.02	-.08
O3	.51	.12	.03	-.04	.02	-.16	.05	.00	.03	.18	.18	.16	.11	.34	.22	-.05	.11	.11
O4	.40	.17	.18	.30	.18	.28	.11	.29	-.04	-.20	-.09	-.09	-.20	-.10	-.16	.07	-.02	-.18
O5	.41	-.08	-.15	-.04	.00	-.11	-.02	-.11	.00	.05	-.06	.04	.01	.10	.03	.09	.01	.05
O6	.28	-.04	.11	.16	.12	.26	.01	.15	-.06	-.26	-.11	-.14	-.07	-.11	-.17	-.05	-.07	-.27
O	.53	.17	.11	.17	.04	.17	.14	.20	.06	-.13	-.03	-.01	-.09	.07	-.03	.02	.09	-.08

(continued on next page)

Table 1 (continued)

IPIP-NEO-300 Self-Report																		
ACQ	C4	C5	C6	C7	N1	N2	N3	N4	N5	N6	N7	O1	O2	O3	O4	O5	O6	O7
E1	.03	.17	.02	.13	-.18	-.23	-.26	-.30	-.06	-.18	-.27	-.07	-.04	.05	.18	-.05	-.05	.00
E2	-.03	.08	-.11	.01	-.15	-.11	-.21	-.35	.10	-.16	-.20	.03	-.05	.00	.18	-.01	-.08	.01
E3	.14	.14	-.01	.10	-.13	.08	-.09	-.37	.00	-.20	-.15	-.02	.06	.12	.21	.13	-.02	.12
E4	.21	.23	.07	.18	-.08	-.05	-.19	-.21	-.15	-.16	-.18	-.10	-.03	.04	.16	.04	.02	.03
E5	-.02	.02	-.29	-.14	-.20	.02	-.11	-.36	.19	-.18	-.14	.12	-.14	-.08	.26	.07	-.02	.05
E6	.01	.18	.10	.14	-.17	-.25	-.27	-.18	-.11	-.13	-.25	.03	.02	.11	.18	.08	.00	.10
E	.04	.12	-.08	.04	-.15	-.09	-.21	-.35	.04	-.16	-.20	-.03	.01	.10	.22	.04	-.04	.07
A1	.01	.13	.07	.14	-.08	-.21	-.19	-.07	.01	-.06	-.14	-.01	.02	.05	.05	-.05	-.04	.00
A2	.01	.05	.09	.12	-.01	-.13	-.12	-.05	-.16	-.06	-.12	.07	.06	.07	-.04	.08	-.01	.06
A3	.05	.13	.16	.18	.05	-.14	-.05	.08	-.12	.01	-.04	-.01	.07	.09	-.06	.01	.00	.03
A4	.03	.09	.14	.18	.10	-.14	.01	.16	-.05	.04	.02	.01	.03	.14	-.12	.03	.05	.04
A5	-.08	.08	.10	.09	.01	-.16	.02	.05	-.11	-.01	-.04	-.02	-.05	.01	-.06	-.09	.01	-.06
A6	.05	.11	.22	.20	.24	-.03	-.01	.23	-.02	.22	.13	.00	.23	.32	-.03	-.11	-.06	.08
A	.02	.14	.18	.19	.04	-.19	-.07	.11	-.07	.03	-.04	.04	.09	.20	.04	.07	.07	.13
C1	.24	.25	.17	.27	-.12	-.09	-.24	-.09	-.09	-.18	-.18	.07	-.08	-.03	.18	.13	.02	.07
C2	.34	.44	.30	.48	.08	-.08	-.08	.10	-.07	.09	.00	.08	.09	.12	-.02	.08	.04	.10
C3	.21	.30	.25	.35	.07	-.13	-.09	.10	-.15	.01	-.05	.03	.01	.08	-.06	-.03	-.03	.00
C4	.37	.35	.27	.39	-.05	-.15	-.20	.01	-.15	-.13	-.15	.00	-.04	-.04	.06	.09	.02	.02
C5	.31	.37	.28	.40	-.03	-.24	-.19	-.02	-.23	-.16	-.19	-.09	-.02	-.02	.09	.07	-.05	-.01
C6	.17	.18	.28	.31	.06	-.19	-.10	.15	-.15	.02	-.05	-.01	.08	.05	-.06	.02	-.02	.02
C	.27	.32	.30	.42	.06	-.11	-.12	.09	-.16	-.06	-.07	-.01	-.03	.04	-.02	.04	-.02	.00
N1	-.16	-.07	-.01	-.10	.44	.29	.40	.36	-.03	.44	.42	-.04	.06	.28	-.23	-.05	.07	.02
N2	-.14	-.17	-.22	-.21	.24	.55	.31	.03	.14	.24	.34	.03	-.03	.10	-.16	-.05	-.01	-.03
N3	-.17	-.29	-.12	-.22	.42	.38	.61	.35	.11	.44	.52	.10	.12	.24	-.24	.04	.06	.08
N4	-.21	-.21	.01	-.16	.46	.22	.44	.47	.04	.46	.46	.08	.14	.30	-.26	.02	.18	.12
N5	-.24	-.32	-.34	-.38	.27	.37	.38	.13	.33	.32	.40	.17	.08	.22	-.07	-.02	-.03	.09
N6	-.19	-.23	-.09	-.19	.43	.27	.46	.34	.11	.44	.46	.09	.26	.31	-.03	.00	.08	.18
N	-.22	-.26	-.10	-.24	.45	.39	.52	.39	.13	.46	.52	.10	.09	.25	-.26	-.03	.05	.05
O1	-.17	-.11	-.07	-.14	.03	-.01	.10	-.03	.08	.04	.05	.26	.08	.18	.10	.18	.21	.26
O2	-.01	-.03	.01	-.01	.17	.11	.13	.08	.03	.11	.14	.23	.36	.31	.12	.19	.20	.37
O3	.00	.03	.10	.07	.35	.11	.09	.19	.09	.26	.23	.09	.31	.42	-.01	.02	.12	.24
O4	.01	.04	-.22	-.07	-.13	-.03	-.01	-.32	.12	-.14	-.11	.07	.11	.07	.28	.16	.08	.20
O5	.09	.06	.15	.10	.06	-.05	-.04	.09	-.08	-.01	-.01	.02	.07	.05	.05	.24	.24	.18
O6	-.01	-.01	-.20	-.14	.00	.12	.12	-.12	.07	.01	.05	.23	.09	.12	.25	.29	.34	.34
O	.07	.04	-.10	.01	.00	-.01	-.02	-.14	.08	-.02	-.03	.15	.19	.20	.21	.20	.21	.30

Note. ACQ = Acquaintance Rating Scales. ICC = Intraclass correlation reliability estimate. Correlations between acquaintance ratings and IPIP-NEO-300 scales for the same construct are in **boldface**. E1 = Friendliness, E2 = Gregariousness, E3 = Assertiveness, E4 = Activity Level, E5 = Excitement-Seeking, E6 = Cheerfulness, E = Extraversion, A1 = Trust, A2 = Morality, A3 = Altruism, A4 = Cooperation, A5 = Modesty, A6 = Sympathy, A = Agreeableness, C1 = Self-Efficacy, C2 = Orderliness, C3 = Dutifulness, C4 = Achievement-Striving, C5 = Self-Discipline, C6 = Cautiousness C = Conscientiousness, N1 = Anxiety, N2 = Anger, N3 = Depression, N4 = Self-Consciousness, N5 = Immorality, N6 = Vulnerability, N = Neuroticism, O1 = Imagination, O2 = Artistic Interests, O3 = Emotionality, O4 = Adventurousness, O5 = Intellect, O6 = Liberalism, O = Openness to Experience.

Table 2
Comparison of percent accurately classified by .5 SD and ODA Methods.

Trait	Overall PAC ^a		ODA Cutting Scores ^b	
	± .5 SD	ODA	Low	High
Friendliness	40	62	−1.35	1.01
Gregariousness	43	77	−1.98	1.78
Assertiveness	48	72	−1.59	.91
Activity Level	41	64	−1.71	1.13
Excitement-Seeking	51	68	−1.88	1.38
Cheerfulness	39	67	−1.57	1.61
EXTRAVERSION	44	70	−2.40	.88
Trust	41	52	−2.03	.64
Morality	36	44	−.14	1.55
Altruism	37	41	−.20	1.41
Cooperation	35	68	−1.03	1.48
Modesty	40	78	−1.89	2.21
Sympathy	39	56	−1.30	.69
AGREEABLENESS	41	33	.17	.78
Self-Efficacy	39	33	−.55	−.03
Orderliness	46	71	−1.79	.94
Dutifulness	39	58	−1.05	.13
Achievement-Striving	44	32	.11	.56
Self-Discipline	37	78	−.59	.88
Cautiousness	35	76	−1.84	2.03
CONSCIENTIOUSNESS	43	53	−.57	1.32
Anxiety	47	66	−.49	1.65
Anger	45	58	−1.41	1.63
Depression	41	67	.01	2.11
Self-Consciousness	41	64	−.11	1.84
Immoderation	41	68	−1.06	2.24
Vulnerability	36	66	−.86	2.31
NEUROTICISM	43	71	−.82	1.91
Imagination	35	77	−2.27	1.38
Artistic Interests	43	68	−2.37	.89
Emotionality	37	63	−1.66	.95
Adventurousness	41	65	−1.51	.62
Intellect	40	57	−1.65	.77
Liberalism	42	79	−1.84	1.88
OPENNESS	44	55	−1.23	.98

^a Percent accurately classified, decimal points omitted.

^b Optimal cutting scores from Optimal Data Analyses, in standard deviations.

Thus, the average ODA low and high cutting scores are closer to what Hofstee and de Raad (1992) noted as the 1.4 SD difference typically required for statistical significance than to the traditional ± .5 SD cutting score.

4. Discussion

Cutting scores from Optimal Data Analysis (ODA; Yarnold & Soltysik, 2005) appear to classify acquaintance perceptions of *low*, *average* and *high* levels of traits more accurately than the convention of classifying scores within ± .5 SD of the mean as *average* and scores outside that range as *low* or *high*. Optimal cutting scores varied considerably across scales, averaging more than 1 SD from the mean.

A number of cautions should be raised about interpreting superiority for the ODA over the traditional ± .5 SD cutting score method from the current data. First, the skewness of the acquaintance ratings could suggest a leniency or positivity "bias" on the part of the acquaintances in the study that might not be found in a random sample from the total population. "Bias" is placed in quotes to indicate that the positivity is not necessarily error. The acquaintance ratings did, after all, correlate appropriately with the self-report scores. Also, in their review of congruence of self- and others' judgments of personality, Funder and Colvin (1997) argue that the generally positive personality descriptions from the acquaintances of a university sample probably reflect reality. The majority of students, they say, possess enough favorable character traits to be admitted to the university. Furthermore, human subjects regulations require participation to be completely voluntary, so only the most cooperative students will participate in a

study such as this one. It also stands to reason that friends and family who agree to provide acquaintance ratings are going to have a generally positive view of the students they are rating. Finally, the fact that the acquaintances seemed willing to rate the students more positively than the scores on the IPIP-NEO-300 might have suggested is striking, given that self-ratings are usually more positive than acquaintance ratings when such ratings are gathered on the same scale (Funder & Colvin, 1997).

Nevertheless, even though the positively skewed acquaintance ratings probably reflect reality for the current convenience sample, it remains unknown how well the ± .5 SD cutting scores might have worked in a study using a genuinely random sample where the acquaintances did not necessarily have a favorable impression of the persons they were rating. Such a study would be difficult to design. A prospective acquaintance judge of personality must know the person he or she is judging for a sufficiently long period of time to provide a valid personality assessment. But, for the most part, people who know each other for long periods of time are people who like each other well enough to stay in each other's lives. Hence, one is likely to get positive personality evaluations from most qualified acquaintances. An exception might be a workplace setting or a live-in setting such as a dormitory where people interact often enough to know each other but are not necessarily good friends. Future research with such a sample can evaluate whether the present finds are replicable under those conditions.

Another caution concerns the replicability of the specific cutting scores computed by ODA. Although ODA cutting scores are derived by the LOO method of cross-validation within the sample, it remains to be seen how well these cutting scores will cross-validate to a new sample.

Yet another caution about the current study concerns the novel, previously untested scales used for gathering acquaintance judgments of personality. The percentile response options from 1 to 99, anchored by the written descriptions *very low* through *very high* might seem, prima facie, like a reasonable format. And all response options, 1 through 99, were used by at least some acquaintances on many of the traits judged. Nonetheless, this is the first time this set of rating scales has been used, and subsequently this is the first information we have about its reliability and validity. The ICC estimates of reliability were not particularly high, which is not surprising given the psychological distance between the researcher and the acquaintance judges. The study design provided no opportunity to explain how to use the scales beyond the instructions given to the acquaintances, and there was no opportunity to train and calibrate the judges, which would have improved reliability. Apparently, the averaged judgments were reliable enough to support statistically significant correlations with the corresponding self-report scales on the IPIP-NEO-300. Nonetheless, clearer results might be seen with differently designed rating scales and/or different procedures to train acquaintance judges.

In the context of these cautions, the current data do suggest that the traditional ± .5 SD cutting scores for self-report scales are probably not optimal for predicting whether acquaintances will perceive respondents as *low*, *average*, or *high* on a particular personality trait. Future studies with more diverse samples are needed to clarify the relative sensitivity, predictive value, effect strengths, and model efficiency for the traditional ± .5 SD cutting score method and the ODA method.

References

- Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist*, 61, 27–41. <https://doi.org/10.1037/0003-066X.61.1.27>.
- Block, J. (1978). *The Q-sort method in personality assessment and psychological research*. Palo Alto, CA: Consulting Psychologists Press.
- Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin*, 136, 1092–1122. <https://doi.org/10.1037/a0021212>.
- Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO personality inventory (NEO PI-R) and NEO five-factor inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.
- Dawes, R. M., & Meehl, P. E. (1966). Mixed group validation: A method for determining

- the validity of diagnostic signs without using criterion groups. *Psychological Bulletin*, 66, 63–67. <https://doi.org/10.1037/h0023584>.
- Funder, D. C., & Colvin, C. R. (1997). Congruence of others' and self-judgments of personality. In R. Hogan, J. A. Johnson, & S. R. Briggs (Eds.), *Handbook of personality psychology* (pp. 617–647). San Diego, CA: Academic Press.
- Goldberg, L. R. (1999). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Vol. Eds.), *Personality psychology in Europe: 7*, (pp. 7–28). Tilburg, the Netherlands: Tilburg University Press.
- Gough, H. G., & Bradley, P. (1996). *CPI manual (third ed.)*. CA: Consulting Psychologists Press Palo Alto.
- Hofstee, W. K. B. (1994). Who should own the definition of personality. *European Journal of Personality*, 8, 149–162. <https://doi.org/10.1002/per.2410080302>.
- Hofstee, W. K. B. (2009). Restyling personality assessments. In L. B. Palfroft, & M. V. Lopez (Eds.), *Personality assessment: New research* (pp. 223–235). Hauppauge, NY: Nova Science Publishers.
- Hofstee, W. K. B., & de Raad, B. (1992). Personality structure through traits. In G.-V. Caprara, & G. L. Van Heck (Eds.), *Modern personality psychology: Critical reviews and new directions* (pp. 56–72). New York: Harvester Wheatsheaf.
- Hofstee, W. K. B., & Hendriks, A. A. J. (1998). The use of scores anchored at the scale midpoint in reporting individuals' traits. *European Journal of Personality*, 12, 219–228. [https://doi.org/10.1002/\(SICI\)1099-0984\(199805/06\)12:3<219::AID-PER315>3.0.CO;2-Y](https://doi.org/10.1002/(SICI)1099-0984(199805/06)12:3<219::AID-PER315>3.0.CO;2-Y).
- Hogan, R., & Hogan, J. (1992). *Hogan Personality Inventory manual*. Tulsa, OK: Hogan Assessment Systems.
- Johnson, J. A. (2005). Ascertaining the validity of web-based personality inventories. *Journal of Research in Personality*, 39, 103–129. <https://doi.org/10.1016/j.jrp.2004.09.009>.
- Johnson, J. A. (2014). Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality*, 51, 78–89. <https://doi.org/10.1016/j.jrp.2014.05.003>.
- Kolar, D. W., Funder, D. C., & Colvin, C. R. (1996). Comparing the accuracy of personality judgments by the self and knowledgeable others. *Journal of Personality*, 64, 311–337. <https://doi.org/10.1111/j.1467-6494.1996.tb00513.x>.
- Rorer, L. G., Hoffman, P. J., & Hsieh, K. C. (1966). Utilities as base-rate multipliers in the determination of optimum cutting scores for the discrimination of groups of unequal size and variance. *Journal of Applied Psychology*, 50, 354–368. <https://doi.org/10.1037/h0023730>.
- Staff, I.P.A.T. (1973). *16 PF test profile sheet*. Institute for Personality and Ability Testing, Champaign, IL.
- Staff, I.P.A.T. (1979). *Administrators manual for the 16 PF*. Institute for Personality and Ability Testing, Champaign, IL.
- Vazire, S. (2006). Informant reports: A cheap, fast, and easy method for personality assessment. *Journal of Research in Personality*, 40, 472–481. <https://doi.org/10.1016/j.jrp.2005.03.003>.
- Yarnold, P. R., & Soltysik, R. C. (2005). *Optimal data analysis. A guidebook with software for windows*. Washington, DC: American Psychological Association.