Journal of Research in Personality 57 (2015) 61–71

Contents lists available at ScienceDirect

Journal of Research in Personality

journal homepage: www.elsevier.com/locate/jrp

A model of "substance" and "evaluation" in person judgments

Daniel Leising [a,*], Stefan Scherbaum [a], Kenneth D. Locke [b], Johannes Zimmermann [c]

[a] Technische Universität Dresden, Germany
[b] University of Idaho, ID, USA
[c] Universität Kassel, Germany

ARTICLE INFO

Article history:
Available online 11 April 2015

Keywords:
Social desirability
Interpersonal
Person perception
Attitude

ABSTRACT

For decades, person perception research has grappled with the distinction between the targets' actual characteristics ("substance") and how positively or negatively those characteristics are viewed by perceivers ("evaluation"); however, lack of an overarching theoretical framework makes it difficult to establish connections between related lines of research. We review the relevant literature, and present and test an algebraic model that incorporates the major insights from that literature. The model posits that all person judgments reflect substance and evaluation to different extents. The evaluation component reflects an interaction between the item's evaluative tone and the perceiver's evaluative attitude regarding the target person. The model may function as an integrative framework that helps improve conceptual clarity and cumulative progress in person perception research.

1. Introduction

Psychological research into the "substantive" versus the "evaluative" components of individual's judgments of themselves and others has been ongoing for decades, and continues to this day. McCrae and Costa (1983) used the terms "substance" and "style" to distinguish between *what is actually there* (substance) and *how it is seen or presented* (style). In their nomenclature, style is about the positive or negative "spin" that the person who delivers the judgment gives to the naked facts (substance) about the target's personality. Presumably, most person descriptions incorporate both components to some extent. However, in this paper we will use the term "evaluation" instead of "style" because the latter term also encompasses influences on person perception that, while independent of substance, go beyond broad positive or negative connotations (e.g., a given perceiver's tendency to see people as being *trustworthy* in particular). The present paper focuses exclusively on those kinds of style that cast a *positive* or *negative* light on targets.

Despite the long history of person perception research, there is still no comprehensive theoretical framework that incorporates the distinction between substance and evaluation. This is very unfortunate, because it makes it difficult to directly compare the research paradigms and findings from different studies. In the present paper, we outline such a framework. As our goal is maximum exactness and parsimony, we present the model in an algebraic format. Using an empirical data set, we then test some of the model's core predictions, and draw the reader's attention to some of its most important implications (e.g., that inter-rater agreement and scale consistency may be high even if the items that are used do not reflect any actual target characteristics at all).

Most research in this field has been based on the (tacit) notion that substance and evaluation are independent of each other; that is, the very same "facts" about a person may be presented with very different evaluative "tones" (Peabody, 1967; Saucier, 1994). Although it is seldom explicitly acknowledged, the basic philosophical position seems to be, first, that there are characteristics that actually exist in the targets, irrespective of the perceiver. These are the substance. And, second, each perceiver may have a more lenient or stern view of the targets, irrespective of their actual characteristics. This is evaluation. Throughout the paper, we will adopt the assumption that substance and evaluation are independent of one another; however, in the discussion section we will broaden the scope somewhat, and address the possibility that substance and evaluation might be related.

The third important factor – apart from the truth that resides in the target and the positive or negative attitude toward the target that resides in the perceiver – is the extent to which a given item reflects the perceiver's attitude. According to our knowledge, Edwards (1953) was the first to introduce the idea that items of personality questionnaires differ systematically in how much of a positive or negative evaluation of a target person they entail when

* Corresponding author.
E-mail addresses: daniel.leising@tu-dresden.de (D. Leising), stefan.scherbaum@tu-dresden.de (S. Scherbaum), klocke@uidaho.edu (K.D. Locke), johannes.zimmermann@uni-kassel.de (J. Zimmermann).

http://dx.doi.org/10.1016/j.jrp.2015.04.002
0092-6566/

endorsed. Edwards showed that ratings of the social desirability of personality questionnaire items correlated very strongly with the average endorsement rates of the items (i.e., the average person attributes positive and does not attribute negative characteristics to himself or herself). Anderson (1968) demonstrated that the social desirability of natural person-descriptive terms could be assessed with very good inter-rater reliability, and that the distribution of item desirability ratings was bi-modal – that is, most person-descriptive terms from the natural language have a clear positive or negative connotation. This finding was later replicated for a different sample of English terms by Dumas, Johnson, and Lynch (2002) and for a sample of German terms by Leising, Ostrovski, and Borkenau (2012). Therefore, most of the terms that we use to describe ourselves and others in everyday life are evaluative.

It follows that the perceiver's evaluative attitude toward a target should predict whether he or she will use positive or negative terms to describe that target. For example, even though the average person tends to endorse positive and not endorse negative items in self-descriptions (see above), there are considerable inter-individual differences in that regard. A person's overall self-evaluations may range from very negative to very positive (Bono & Judge, 2003; Furr & Funder, 1998; Judge, Erez, Bono, & Thoresen, 2002; Leising et al., 2013), and the evaluative tones of items should determine the extent to which such differences influence the person's self-ratings. In other words, the evaluative tone of an item may be expected to *interact* with the perceiver's evaluative attitude regarding the target (e.g., him- or herself) (Bäckström, Björklund, & Larsson, 2009; Leising & Borkenau, 2011; McCrae & Costa, 1983). This interaction between the perceiver's evaluative attitude regarding the target and the evaluative tone of the item is one of the key ingredients of the model that we present below. Note that the same logic applies to self-ratings and other-ratings: A perceiver who has a positive (negative) attitude toward another person should endorse positive (negative) items in describing that person, and the same should be true of a person who likes (dislikes) and describes himself or herself. In the latter case, the perceiver simply happens to be identical with the target.

It is important to note that, when we speak of the perceiver's evaluative attitude, we mean an attitude that already existed before the current judgment takes place. This needs to be emphasized in order to clarify that we do not address impression formation regarding *new* targets about which the perceiver knows nothing yet. A whole body of literature deals with the formation of (e.g.) approach and avoidance tendencies regarding new stimuli, but this is not our subject in the present paper. Rather, we deal with descriptions of targets toward whom the perceiver already *has* a firmly established evaluative attitude.

A series of studies recently corroborated the notion of an interaction between evaluative item and perceiver characteristics in person judgments. Borkenau and Zaltauskas (2009) found that participants with higher *self-esteem* (i.e., participants who liked themselves more) described themselves in more "normative" ways, that is, they endorsed items to the extent that they were endorsed by the participant sample on average. Because such normative ratings tend to be extremely similar to "ideal" ratings (Edwards, 1953), the participants' responses could be predicted by an interaction of how much they liked themselves (self-esteem, evaluative attitude) and evaluative item tone. Leising, Erbs, and Fritz (2010) showed that rated item desirability moderates the extent to which *other*-ratings of personality may be predicted from how much the perceivers like the targets: The extent to which the perceivers' liking for the targets predicts the perceivers' ratings of the targets on a personality item strongly depends on how much the item entails a positive or negative evaluation. Notably, such associations persist when perceivers with different evaluative attitudes describe the exact *same* targets (Leising, Ostrovski, & Zimmermann, 2013), and even when perceivers with different evaluative attitudes describe the exact same *behaviors* of the exact same targets (Leising, Gallrein, & Dufner, 2014). These studies provide support for the notion that the perceivers' evaluative attitudes influence person judgments at least partly independent of the targets' actual characteristics.

Notably, if judgments of targets by perceivers are affected by an interaction between the perceivers' evaluative attitudes and the items' evaluative tones, then *correlations* between judgments should be affected by these factors as well. One kind of association between judgments that might be affected is *inter-rater agreement*. In a seminal study, John and Robins (1993) hypothesized that more evaluative (i.e., positive *or* negative) items should yield lower inter-rater agreement, and provided evidence for such a curvilinear relationship using two data sets. A few studies corroborated this relationship, but other studies did not. A recent meta-analysis (Kenny & West, 2010) did *not* support the general notion of a relationship between item evaluativeness and inter-rater agreement. Clearly, that relationship is moderated by other factors that are yet ill-understood. In the present study, we aim to clarify this issue a bit more.

What is most important in the present context, however, is the *reasoning* behind the John and Robins (1993) study: Although it is not made very explicit in their paper, the expectation that more evaluative items should yield lower inter-rater agreement seems to reflect the notion that responses to more evaluative items are more likely to reflect differences in the evaluative attitudes that the different perceivers have toward the targets (e.g., themselves), and that such differences introduce error variance in computing inter-rater agreement. This is essentially the idea of an interaction between item tone and perceiver attitude that we discussed above, although it was not directly tested in the study by John and Robins (1993). Our model incorporates that idea, but clarifies that inter-rater agreement may actually be *higher* for more evaluative items, if the perceivers' evaluative attitudes toward the individual targets are shared.

Inter-rater agreement is usually tested by correlating assessments of the *same* targets on the *same* item by *different* perceivers. However, the interaction between perceiver attitude and item tone may also be expected to affect correlations between assessments of the same targets by the *same* perceivers on *different* items; we will refer to this latter case as "internal consistency". If responses to evaluative items partly reflect the perceivers' evaluative attitudes toward the targets, then different evaluative items that are completed by the *same* perceivers regarding the *same* targets should reflect the *same* evaluative attitudes and thus correlate more strongly with each other (i.e., the correlation should deviate more from zero) than more evaluatively neutral items. In other words, the perceivers' evaluative attitudes toward the targets could be conceived of as a common source of variance (= a factor), that affects responses to items to the extent that the items are evaluative (McCrae & Costa, 1983). Bäckström et al. (2009) provided evidence in favor of this notion, by showing that the first factor in responses to a set of self-report items could be substantially weakened by simply phrasing items less evaluatively. A likely explanation for this finding is that the perceivers' self-ratings on the neutralized items reflected their (unchanged) evaluative attitudes toward themselves to a lesser degree. As a result, the correlations between the items reflected this common source of variance to a lesser degree, leading to a reduced Eigenvalue in the factor analysis. Our model integrates this line of reasoning as well.

1.1. The model

We will now attempt to formulate an integrative algebraic framework that incorporates the crucial insights from the

literature reviewed above. To make the presentation as easy to follow as possible, we focus on item-wise analyses (as opposed to profile analyses; cf. Furr, 2008), as they represent the most common approach in person perception research to date. Thus, we will introduce the model in terms of ratings of many targets (t) by one perceiver (p) on one item (i).

The basic assumption of the model is that the scores Y_{pti} that a perceiver (p) assigns to targets (t) on an item (i) reflect three influences, two of which are systematic, and one of which is unsystematic. The first systematic component reflects some "substantive" quality of the *targets*, that is, something that is really there, irrespective of who does the judgment. We will denote the targets' true scores T_{ti}, and their weight (the extent to which item i reflects individuals' true scores) t_i. The second systematic component reflects the *perceiver's* evaluations of the targets, irrespective of the targets' actual personalities. It is represented by the product of the perceiver's evaluative attitudes regarding the individual targets A_{pt} and a weight a_i that reflects the extent to which item i reflects perceivers' evaluative attitudes. In line with research findings reported above, we assume that this weight is more or less identical with the evaluative tone of the item. Thus, the perceiver's responses to the item Y_{pti} may reflect the perceiver's evaluative attitudes toward the targets A_{pt} only if the item entails at least some positive or negative evaluation ($a_i <> 0$). For the time being, we will assume that perceivers do not differ in how they use items for expressing their evaluative attitudes toward targets. In the Discussion, however, we will briefly address the possibility that the evaluative tones of items may be different for different perceivers. The third component in the model is measurement error E_{pti}. The ratings of a set of targets by a perceiver on an item may thus be decomposed as follows:

$$Y_{pti} = t_i \cdot T_{ti} + a_i \cdot A_{pt} + E_{pti}$$

Obviously, our model equals the basic formula of classical test theory, complemented by the interaction between perceiver attitudes A_{pt} and item tone a_i, plus a weight t_i for the true scores. Introducing the latter has the important conceptual implication that some items may measure *no* real quality of targets at all ($t_i = 0$). It should be noted that, other than in classical test theory, we do not assume that the targets' true scores may only be obtained by averaging across an infinite number of repeated assessments (assuming no evaluation for the time being). Rather, our model is more general than that, because true scores may be defined at will. For example, a researcher may decide that the results of some intelligence test constitute a suitable true score variable for modeling individuals' ratings of one another on the item "smart". In this case, t_i would simply be the extent to which the ratings reflect the test scores, as measured by (e.g.) a multiple regression weight (cf. West & Kenny, 2011).

Our model makes explicit how items may reflect both (a) high or low levels of an actual trait and (b) positive or negative "presentations" of that trait. This is exactly the distinction motivating Peabody's (1967) work using sets of four items to capture the four possible combinations of trait level and evaluation. For example, consider the following set of four adjectives (cf. Borkenau & Ostendorf, 1989): The adjective pair "firm" and "severe" is similar in regard to substance, but dissimilar in regard to evaluation. In terms our model, if "firm" and "severe" are items 1 and 2, then $t_1 = t_2$ and $a_1 \neq a_2$. That is, they measure (more or less) the same actual quality of the targets, but item 1 ("firm") measures it with a positive connotation, whereas item 2 ("severe") measures it with a negative connotation. The same applies to the adjective pair "lenient" and "lax". Thus, there is descriptive consistency, but an evaluative contrast within each of these adjective pairs. On the other hand, the adjective pair "firm" and "lenient" is similar in regard to evaluation ($a_1 = a_2$), but differs in regard to substance ($t_1 \neq t_2$), and the same applies to the adjective pair "severe" and "lax". Peabody (1967), as well as many researchers after him (e.g., Borkenau & Ostendorf, 1989; Locke & Christensen, 2007; Locke, Craig, Baik, & Gohil, 2012; Locke, Zheng, & Smith, 2014) used such adjective pairs to assess individuals' tendencies to portray themselves positively (or negatively) at the cost of providing descriptively inconsistent self-portrayals at the same time. Note that whereas in previous work the truth and attitude weights were implicitly assumed to be dichotomous (high/low), they may vary continuously in our model.

In the following, we will concentrate on how the model may be applied to *associations* between measurements. Table 1 displays the data structure that is relevant for the subsequent presentation. The most basic case (re-test) in which a measurement (Y_{pti}) is simply correlated with another (later) measurement of the exact same kind ($Y_{pti'}$) will not be considered further. Rather, we will concentrate on two cases of particular theoretical and practical importance. The first of these is *internal consistency*, where the same targets are assessed by the same perceiver on two different items. For example, Prudence (perceiver 1) may judge a set of targets (t) in terms of how intelligent (item 1) and how jovial (item 2) they are. In Table 1, the correlation between the measurements in rows 1 and 3, and the correlation between the measurements in rows 2 and 4, reflect the case of internal consistency. The second case is *inter-rater agreement*, where the same targets are judged on the same item by two different perceivers. For example, Prudence (perceiver 1) and Paul (perceiver 2) may judge a set of targets (t) in terms of how intelligent (item 1) they are. In Table 1, the correlation between the measurements in rows 1 and 2, and the correlation between the measurements in rows 3 and 4, reflect the case of inter-rater agreement.

The model formula will now be applied to these different kinds of correlations between measures. The crucial issue is that one core source of variance (T_{ti}) is between-targets whereas another core source of variance (A_{pt}) is between-perceivers, and a third core source of variance (t_i, a_i) is between-items. This leads to specific predictions regarding the different kinds of associations between measures, depending first and foremost on which sources of variance the different measures share. In our empirical analyses, the targets are always the same persons, because we use so called "item-wise" correlations (see below). It may still matter, however, whether the perceivers or the items are the same or not. As the

Table 1
Data structure.

Perceiver (p)	Targets (t)	Item (i)	Y_{pti}	t_i	T_{ti}	a_i	A_{pt}	E_{pti}
1	1	1	Y_{111}	t_1	T_{11}	a_1	A_{11}	E_{111}
	2		Y_{121}	t_1	T_{21}	a_1	A_{12}	E_{121}
	...		...	...	...	...	...	...
	n		Y_{1n1}	t_1	T_{n1}	a_1	A_{1n}	E_{1n1}
2	1	1	Y_{211}	t_1	T_{11}	a_1	A_{21}	E_{211}
	2		Y_{221}	t_1	T_{21}	a_1	A_{22}	E_{221}
	...		...	...	...	...	...	...
	n		Y_{2n1}	t_1	T_{n1}	a_1	A_{2n}	E_{2n1}
1	1	2	Y_{112}	t_2	T_{12}	a_2	A_{11}	E_{112}
	2		Y_{122}	t_2	T_{22}	a_2	A_{12}	E_{122}
	...		...	...	...	...	...	...
	n		Y_{1n2}	t_2	T_{n2}	a_2	A_{1n}	E_{1n2}
2	1	2	Y_{212}	t_2	T_{12}	a_2	A_{21}	E_{212}
	2		Y_{222}	t_2	T_{22}	a_2	A_{22}	E_{222}
	...		...	...	...	...	...	...
	n		Y_{2n2}	t_2	T_{n2}	a_2	A_{2n}	E_{2n2}

Note: Y_{pti} = Perceiver p's description of the targets on item i. t_i = Weight of the targets' true scores. T_{ti} = Targets' true scores. a_i = Weight of the perceiver's evaluative attitudes. A_{pt} = Perceivers' evaluative attitudes. E_{pti} = Measurement error.

most novel component of our model concerns the interplay of the perceivers' evaluative attitudes toward the targets (A_{pt}) and the evaluative tone of the item (a_i), we will focus on the influence of these variables, *independent of substance*: For example, we ask whether measures may correlate strongly with each other (between perceivers, or between items), even though they do not assess the same actual qualities of the targets.

2. Testing the model

We tested the model using a data set from a study by Leising et al. (2013). In contrast to that study, which analyzed "profile correlations" (cf. Furr, 2008), we used the data for analyses of "item-wise correlations". That is, the ratings of all perceivers on all items concerned the same set of target persons. In the Leising et al. (2013) study, the perceivers were asked to report how much they *liked* each target. This variable (Liking) was used to operationalize the perceivers' evaluative attitudes toward the targets (A_{pt}) in the present study.

2.1. Sample

A convenience sample of 209 research volunteers (120 female, 87 male, 2 unreported) participated in the study. Their mean age was 23.1 years (SD = 4.2). Most participants were university students; accordingly, the average level of education in the sample was high.

2.2. Procedure

The participants were asked to describe each of 15 target persons. The 15 target persons were public figures. They had been selected from a larger pool of 50 targets according to the criteria of (a) being known to many people, (b) differing in their apparent personalities, and (c) evoking a large within-target variance in Liking between perceivers. We used public figures as targets in order to be able to obtain a large sample of perceivers who could judge them. By using the same sample of targets for all judgments, we intended to keep the "substance" component constant across judgments: If all judgments refer to the same set of targets, differences between judgments (of the same targets on the same items) may not be attributed to actual personality differences between the targets. Given that the targets in the present study were public figures, we assumed that the perceivers would have access to largely the same information about the targets.

Ratings were provided in an online format. The participants used a list of 30 adjectives (Borkenau & Ostendorf, 1998) that assess the Big Five personality factors by 6 items each (response options ranging from 1 = "doesn't fit at all" to 5 = "fits perfectly"). We excluded ratings with either more than 2 missing values (i.e., fewer than 28 items used to describe the target) or more than 25 items with equal values (suggesting careless responding). We also excluded perceivers who had judged fewer than 12 of the 15 targets. The perceivers were asked to report how much they liked each of the 15 targets (using a 5-point scale ranging from 1 = "not at all" to 5 = "very much"). This information was used to operationalize the evaluative attitude (A_{pt}) component of our model. Ratings of targets for whom liking was not reported were also excluded. Altogether, we analyzed 2673 ratings of 12 or more targets by 189 perceivers. In a previous study (Leising et al., 2010), the 30 items had been rated for their social desirability with very good inter-rater reliability (ICC(3,24) = .98). We used these ratings in the present study, suspecting that rated desirability would be strongly associated with the empirically determined attitude weights (a_i).

2.3. Application of the basic model formula

For computing the targets' true scores (T_{ti}) on the 30 items, we first averaged – separately for each target – the ratings that the target had received from all perceivers who had reported liking him/her to the same extent. These average ratings were then averaged across the five Liking levels – still separately for each target – which resulted in (15 targets × 30 items =) 450 true scores that were balanced for Liking (i.e., they represented an average of 5 average perceivers with different liking levels, giving each Liking level the same weight).

We simultaneously predicted the observed scores (Y_{pti}) from the target's true scores (T_{ti}) and from the perceivers' attitudes (A_{pt}), using a linear mixed-effect model. In this model, we also included four random coefficients: A random intercept for perceivers (taking into account that perceivers may differ in their tendency to endorse any item ["acquiescence"]), a random intercept for items (taking into account that items may differ in base rates), and two random slopes for the influence of T_{ti} and A_{pt} across items (taking into account that items may differ in how strongly true scores and perceiver attitudes affect the observed scores). The two random slopes are conceptually equivalent to t_i and a_i in our model outlined above, and the estimated values for each item were saved for use in later analyses. All variables were centered at their theoretical mean (i.e., Y_{pti}, T_{ti}, and A_{pt} now varied from −2 to 2).

The results of this analysis ("Model I") are displayed in Table 2. Observed scores were strongly influenced by true scores ($Beta$ = 0.82, p < .001), but not by perceivers' attitudes ($Beta$ = 0.04, p = .36). However, analyses of random slopes indicated that the influence of the perceivers' attitudes varied much more strongly across items ($\sigma^2_{i(A)}$ = 0.053) than did the influence of the targets' true scores ($\sigma^2_{i(T)}$ = 0.014). We suspected that the different extents to which the items reflected the perceivers' attitudes toward the targets should be closely mirrored by ratings of the items' social desirability (cf. Leising et al., 2014), and we provide a test of this hypothesis below. The extents to which the items reflected truth and attitudes (i.e., the two random slope variables) were largely uncorrelated (r = .11) across items, which is in line with the most common conceptualizations of the relationship between substance and evaluation (e.g., McCrae & Costa, 1983; Peabody, 1967). In addition, analyses of random intercepts showed that variance in

Table 2
Linear mixed-effect model analysis: Predicting observed scores on an item.

	Model I			Model II		
	Beta	SE	Stand. Beta	Beta	SE	Stand. Beta
Fixed effects						
Intercept	0.060	0.088		0.010	0.031	
T_{ti}	0.815***	0.022	.536	0.813***	0.023	.536
A_{pt}	0.039	0.042	.036	0.015	0.014	.036
$SocDes_i$				0.279***	0.019	.306
$SocDes_i \times T_{ti}$				0.003	0.014	.003
$SocDes_i \times A_{pt}$				0.134***	0.009	.196
Random effects						
σ^2_p	0.016			0.016		
σ^2_i	0.230			0.026		
$\sigma^2_{i(T)}$	0.014			0.014		
$\sigma^2_{i(A)}$	0.053			0.006		
σ^2_e	0.799			0.799		

Note: N = 80,114. T_{ti} = Targets' true scores on the item. A_{pt} = Perceivers' evaluative attitudes toward targets (liking). σ^2_p = Random effect for perceiver (acquiescence). σ^2_i = Random effect for item (base rate). $\sigma^2_{i(T)}$ = Random effect for influence of targets' true scores on observed scores. $\sigma^2_{i(A)}$ = Random effect for influence of perceivers' evaluative attitudes on observed scores. σ^2_e = Error variance. Random covariances are omitted in this table.
*** p < .001.

observed scores was much more due to item differences in base rates ($\sigma^2_i = 0.230$) than to perceiver differences in acquiescence ($\sigma^2_p = 0.016$).

In the next step (“Model II”, see Table 2), we added the social desirability ratings of the 30 items as additional predictors, including their two-way interactions with T_{ti} and A_{pt}. Analyses of fixed effects suggested that observed scores were still strongly predicted by true scores (*Beta* = 0.81, $p < .001$), but also by the social desirability of the item (*Beta* = 0.28, $p < .001$), and an interaction of the item's social desirability and the perceiver's attitude (Beta = 0.13, $p < .001$). The main effect of item desirability implies that the perceivers were inclined to endorse items the more the items had a positive connotation. Note that this effect is not part of our model, and will be ignored in our analyses of item-wise agreement between measures because correlations are unaffected by mean differences between variables.

Fig. 1 displays simple slopes resulting from this analysis. In line with our theoretical model, the attitudes of the perceivers toward the targets only influenced observed scores when the item entailed a positive or negative evaluation. Ceteris paribus, a perceiver who likes a target very much ($A_{pt} = 2$) would be expected to score roughly 1 point higher (assuming a 5-point scale) on a socially desirable item (desirability = 2) than a perceiver who does not like the target at all ($A_{pt} = -2$). For a socially undesirable item (desirability = −2), a perceiver who likes a target very much ($A_{pt} = 2$) would be expected to score roughly one point lower than a perceiver who does not like the target at all ($A_{pt} = -2$). For a neutral item (desirability = 0), perceiver attitudes toward targets would not be expected to affect observed scores.

Notably, whereas perceivers with neutral attitudes ($A_{pt} = 0$) would be expected to assign higher scores with increasing social desirability of the item, (see above), they would not be expected to consider the item's social desirability much when rating targets they do not like ($A_{pt} = -2$). This finding may seem strange at first, because one could imagine that perceivers with negative attitudes tend to ascribe negative, and not to ascribe positive characteristics to their targets. However, that seems not to be the case. Rather, at the lower end of the liking continuum the perceivers' normative assumption that targets have more positive than negative characteristics seems to get lost.

Finally, analyses of random effects indicated that the variance in item intercepts ($\sigma^2_i = 0.026$) and in the influence of perceiver attitudes across items ($\sigma^2_{i(A)} = 0.006$) was reduced by almost 90% when entering ratings of item desirability into the model. This suggests that differences in rated social desirability between items accounted for this variance. In contrast, variance in perceiver intercepts ($\sigma^2_p = 0.016$) and in the influence of true scores across items ($\sigma^2_{i(T)} = 0.014$) remained the same. We also repeated all of these analyses incorporating target-specific liking (averaged across perceivers) as an additional predictor, but the results remained virtually unchanged.

2.4. Predicting correlations between measurements: Internal consistency

We will now discuss the implications of our model for correlations between measurements. Fig. 2 depicts the case of internal consistency. For the sake of simplicity, we will abbreviate the correlation between two measures r_{YY}. In the case of internal consistency, r_{YY} is the association between two ratings of the same targets by the same perceiver on two different items (Y_{pti}, Y_{ptj}). That is, we study how various combinations of two items (e.g., friendly-intelligent; intelligent-lazy; lazy-attractive, etc.) on which the same perceiver describes a set of targets correlate with one another. Note that Fig. 2 only displays the model for one particular pair of items. The overall analysis of how r_{YY} depends on variations in a_i, t_i, A_{pt} and T_{ti} comprises ($189 * 30 * 29/2=$) 82,215 such cases (because each of the 30 items is correlated with every other item, and this is done for each of the 189 perceivers; however, there were some missing values).

As evident from Fig. 2, our model predicts that, in the case of internal consistency, r_{YY} should depend on two influences in particular: One is the three-way interaction between the true score correlation (r_{TT}) between T_{ti} and T_{tj} on the one hand, and the product of t_i and t_j on the other hand. That is, the more *each* of the two items reflects something that is real (t_i, t_j), *and* the more those real

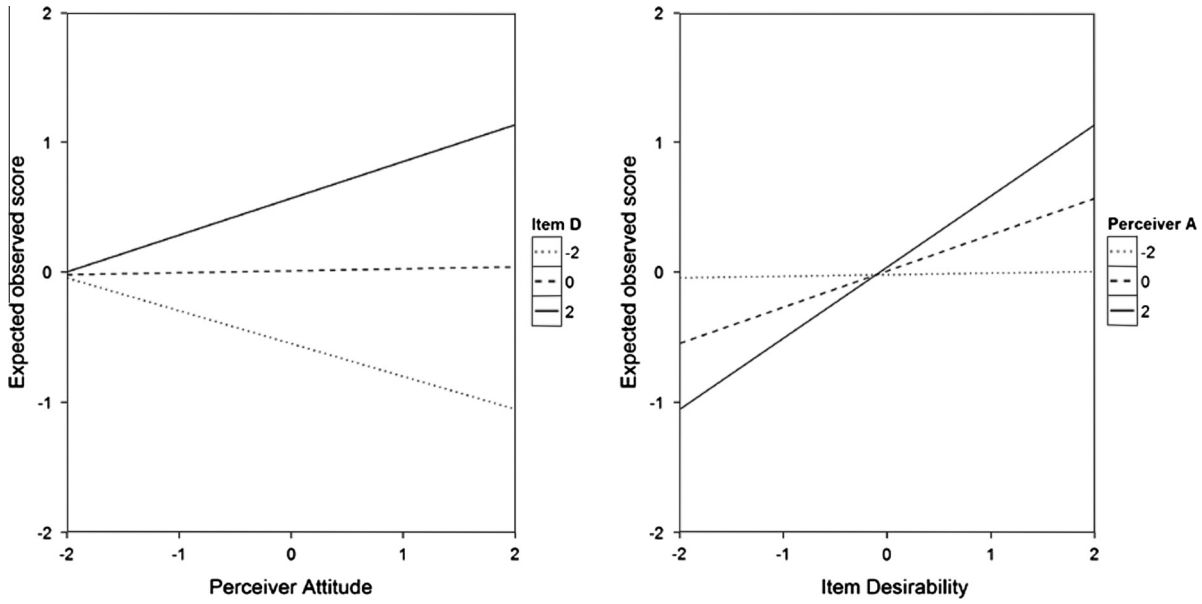

Fig. 1. Simples slopes resulting from the test of the basic model formula: Only evaluative items reflect perceiver attitudes. The left image shows how the effect of evaluative perceiver attitudes (predictor, *x*) on personality ratings (criterion, *y*) is moderated by the social desirability of the respective item (three lines): More positive perceiver attitudes yield higher ratings on socially desirable items, and lower ratings on socially undesirable items. In contrast, perceiver attitudes do not affect observed scores for evaluatively neutral items. Very negative perceiver attitudes do not yield different ratings on items differing in social desirability. The right image shows the exact same effect, with the roles of predictor (item desirability, *x*) and moderator (perceiver attitude) reversed.

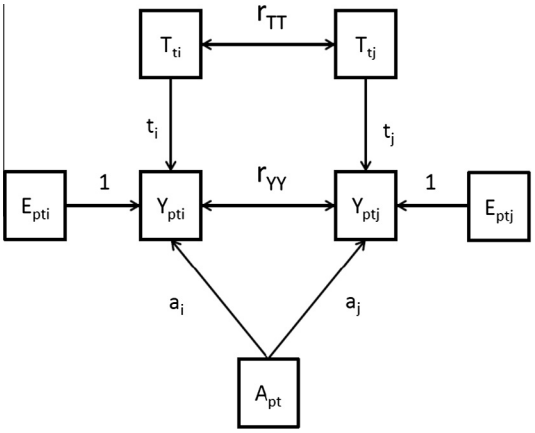

Fig. 2. Model for predicting the internal consistency (r_{YY}) of two items (i, j) that are used by the same perceiver (p) for describing the same targets (t). Apart from measurement error, the correlation r_{YY} between the two measurements (Y_{pti}, Y_{ptj}) should depend on the extent to which *both* items assess the perceiver's evaluative attitudes A_{pt}. This extent is the product of a_i and a_j. The correlation between the two measurements should also depend on the interaction between the correlation r_{TT} of the targets' true scores on the two items (T_{ti}, T_{tj}) and the extent to which *both* items reflect those true scores. The latter is the product of t_i and t_j.

things are the same (r_{TT}), the higher the internal consistency of the two measures (r_{YY}) should be. The other likely influence is the two-way interaction of a_i and a_j: The more the two measures reflect the perceiver's attitudes toward the individual targets in the same way (e.g., positively), the more the internal consistency of the two measures should increase. If one item reflects the perceiver's attitudes positively (e.g., "bright") and the other item reflects the perceiver's attitudes negatively (e.g., "smartass"), the correlation between the two items should decrease and possibly even become negative.

To test whether these two predictions are valid we ran a linear mixed-effect model analysis. Note that we used the "double entry" method, i.e., each combination of items was included twice.[1] We did so in order to account for the fact that each item could arbitrarily be assigned the role of item i or item j within each item pair (i.e., items were indistinguishable; cf. Kenny, Kashy, & Cook, 2006). We predicted the observed correlation r_{YY} from the true score correlation r_{TT}, the truth weights t_i and t_j, their two-way and three-way interactions, as well as the attitude weights a_i and a_j and their two-way interaction. The a_i, a_j, t_i and t_j weights we used in this analysis were the ones that had been estimated as random slopes in the first round of analyses described above. In addition, we included random intercepts for perceivers and items, allowing for the possibility that some perceivers tend to provide more (or less) similar judgments on different items than the average perceiver does, and that items may differ in how well they converge with other items, on average. All variables were centered prior to analysis.

The results of the analysis suggested that the observed correlation between two items depends on two influences in particular, as evidenced by comparisons of standardized effect sizes (see Table 3): The correlation between the targets' true scores (*Beta* = .39) and the interaction between the items' attitude weights (*Beta* = 2.15). The fact that, in this analysis, r_{YY} was predictable from the true score correlation rather than the three-way interaction between t_i, t_j and r_{TT} contradicted our expectations. However, we think that this finding may be explained in terms of (a) how the true scores were obtained in the present study and (b) the resulting empirical range of t_i and t_j. We obtained the targets' true scores by averaging observed scores across perceivers (balanced for liking). Then we used these true scores to predict the observed scores, and thereby obtained the true score weights t_i. In order for r_{TT} to be high, both t_i and t_j thus had to be relatively high as well. As a consequence, the product of t_i and t_j could not add much information beyond the true score correlation r_{TT}. The second possible reason for the unexpected finding in regard to the influence of $t_i * t_j * r_{TT}$ is the relatively low variance of t_i and t_j in our dataset (range: 0.57–1.02), limiting the possibility for the three-way interaction to deviate markedly from the true score correlation r_{TT}. It should be noted, however, that both of these restrictions (a and b) do not have to apply to other datasets. If "true scores" were rationally chosen, rather than computed as averages of the measurements that are to be predicted, r_{TT} could easily be high while t_i, t_j, or both are close to zero (and vice versa). Consider the following two cases (each could be displayed in the format of Fig. 2): In the first hypothetical case, we use the results of a vocabulary test as the true score variable for the item "articulate", and the results of a general intelligence test as the true score variable for the item "beautiful" (!). In this case, the two true score variables would probably be closely associated, but the three way interaction between r_{TT}, t_i, and t_j would be close to zero, because general intelligence is not a very good predictor of rated beauty (i.e., the true score weight for this item would be very low). In the second case, we only replace the item "beautiful" by "smart". This time, the three-way interaction should be much higher. If we used only these two cases in our analysis described above, we would find that the three-way interaction, rather than the true score correlation alone, predicts the observed correlation between two items.

In regard to the factors that contribute to internal consistency, the role of the attitude weight product is particularly remarkable: It implies that we will be able to partly predict the correlation between two items simply from knowing how much they both entail positive or negative evaluations of targets. This effect was relatively strong: For example, if both a_i and a_j equal .30 then the

Table 3
Linear mixed-effect model analysis: Predicting internal consistency (i.e., the correlation between two items by which the same perceiver describes the same set of targets).

	Beta	SE_{jack}	Stand. Beta
Intercept	0.003*	0.001	
r_{TT}	0.387***	0.009	.415
t_i	−0.014*	0.006	−.004
t_j	−0.014*	0.006	−.004
a_i	0.075***	0.007	.044
a_j	0.075***	0.007	.044
$r_{TT} \times t_i$	0.209***	0.032	.025
$r_{TT} \times t_j$	0.209***	0.032	.025
$t_i \times t_j$	−0.092	0.057	−.003
$a_i \times a_j$	2.152***	0.111	.288
$r_{TT} \times t_i \times t_j$	1.672***	0.277	.022

Note: N = 81,562 * 2 (double entry method). r_{TT} = Correlation between the targets' true scores on the two items (i, j). t_i = influence of targets' true scores on observed scores on item i. t_j = influence of targets' true scores on observed scores on item j. a_i = influence of perceivers' evaluative attitudes on observed scores on item i. a_j = influence of perceivers' evaluative attitudes on observed scores on item j. Effects describe influences of (combinations of) variables when all other variables equal their own individual averages (i.e., a = 0.04, t = 0.82, and r_{TT} = −0.02). Standard errors (SE) and p-values are based on the Jackknife method. Random effects are omitted in this table. Standardized betas were computed on the basis of variables that were z-transformed in the long data format.

* $p < .05$.
** $p < .01$.
*** $p < .001$.

[1] In these analyses, we relied on a jackknife resampling procedure to test whether the fixed effects were statistically significant (Efron & Tibshirani, 1993). This was necessary because of numerous non-independencies in the data. Specifically, we computed the respective coefficients (N=) 189 times, each time omitting the data from exactly one of the 189 perceivers. The distribution of coefficients across these 189 resamples was then used to estimate standard errors, which were then used to derive p-values. In a previous publication, we presented evidence from a Monte Carlo simulation study showing that jackknife based p-values are sufficiently accurate for data structures such as ours (Leising et al., 2013).

correlation between the two items would be expected to increase by (0.30 ∗ 0.075 + 0.30 ∗ 0.075 + 0.30 ∗ 0.30 ∗ 2.152=) 0.24, *controlling for substance*. This is highly relevant to the debate over the so-called "general factor of personality" (see below). It should be noted, however, that the product of a_i and a_j may only affect internal consistency if the attitudes of the perceiver regarding the individual targets differ from one another. Table 3 also includes a number of additional significant effects that were not predicted by our model. As these were considerably smaller than the ones discussed above, for simplicity we will ignore them here.

2.5. Predicting correlations between measurements: Inter-rater agreement

In the context of our study, inter-rater agreement was operationalized as the correlation between judgments by two perceivers (p, q) who used the same item (i) for judging the same set of targets (t). For example, Priscilla and Quinn may judge the intelligence of a set of targets. Fig. 3 displays such a constellation in abstract form. Note again that the figure only incorporates two perceivers and one item, whereas the full analysis we present below incorporates hundreds of such constellations. According to our model, we would expect inter-rater agreement (r_{YY}) to be predictable from two influences in particular: The first is the extent to which the item measures some real quality of the targets, t_i^2. The fact that the square of t_i rather than t_i is relevant here implies that inter-rater agreement will improve the more the correlation between the true scores T_{ti} and both of the measured variables (Y_{pti} and Y_{qti}) deviates from zero (in *either* direction). The second relevant predictor is the interaction between the correlation of the perceivers' evaluative attitudes toward the targets (r_{AA}) and the (squared) extent to which the item reflects these attitudes (a_i): The more two perceivers tend to favor *the same* targets over other targets, and the more the item reflects such preferences, the more their agreement in judging the targets on the item should be inflated. The fact that the square of a_i rather than a_i is relevant here implies that it does not matter whether the item has a *positive* or *negative* evaluative connotation. To the extent that the perceivers' evaluative attitudes toward the targets are shared, inter-rater agreement will be improved the more the item is evaluative (i.e., a_i <> 0).

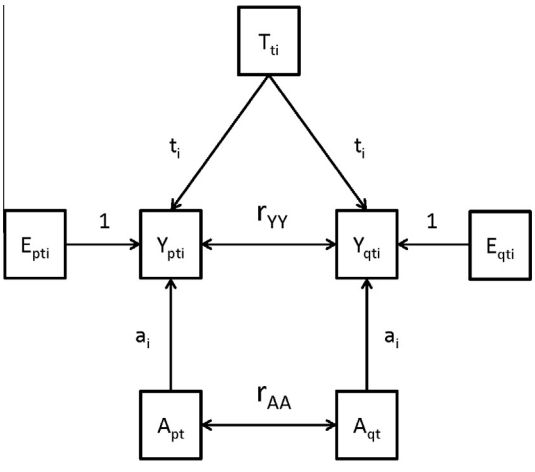

Fig. 3. Model for predicting inter-rater agreement (r_{YY}) between two perceivers (p, q) that use the same item (i) for describing the same targets (t). Apart from measurement error, the correlation r_{YY} between the two measurements (Y_{pti}, Y_{qti}) should depend on the extent t_i^2 to which the item assesses the targets true scores, regardless of the direction of that association. The correlation between the two measurements should also depend on the interaction between the correlation r_{AA} of the perceiver's evaluative attitudes toward the targets (A_{pt}, A_{qt}), and the extent a_i^2 to which the item reflects those attitudes, regardless of the direction of that association.

Table 4
Linear mixed-effect model analysis: Predicting inter-rater agreement (i.e., the correlation between ratings of the same set of targets on an item by two different perceivers).

	Beta	SE_{jack}	Stand. Beta
Intercept	0.224***	0.025	
t_i	0.467***	0.033	.174
t_i^2	−0.626***	0.160	−.026
a_i	0.031*	0.015	.024
r_{AA}	0.007	0.006	.007
a_i^2	0.483***	0.095	.084
$a_i \times r_{AA}$	0.082***	0.020	.019
$a_i^2 \times r_{AA}$	1.050***	0.125	.055

Note: N = 527,592 ∗ 2 (double entry method). r_{AA} = Correlation between the two perceivers' (p, q) evaluative attitudes toward the targets. t_i = influence of targets' true scores on observed scores on item i. a_i = influence of perceivers' evaluative attitudes on observed scores on item i. Effects describe influences of (combinations of) variables when all other variables equal their own individual averages (i.e., a = 0.04, t = 0.82, and r_{ApAq} = 0.12). Standard errors (SE) and p-values are based on the Jackknife method. Random effects are omitted in this table. Standardized betas were computed on the basis of variables that were z-transformed in the long data format.

* p < .05.
** p < .01.
*** p < .001.

We tested our predictions using linear mixed-effect model analysis. For this analysis, we first computed all correlations for judgments of the targets by the (189 ∗ 188/2=) 17,766 pairs of individual perceivers for each of the 30 items. The resulting correlations were entered into the model as the dependent variable, and predicted from variations in the truth weight t_i and its square, as well as the attitude correlation r_{AA}, the attitude weight a_i and its square, and the two-way interactions between attitude correlation and the (squared) attitude weight. Again, we used the double entry method, included random intercepts for perceivers and items, and centered all variables prior to analysis.

The analysis (see Table 4) yielded a significant intercept of .22 which is very close to an estimator of average consensus reported by Kenny (2004; cf. Kenny, Albright, Malloy, & Kashy, 1994). The analysis also yielded a significant standardized Beta of .17 for the weight of the true scores, which was the strongest effect of any of the predictors. Therefore, inter-rater agreement will in fact be higher the more the respective item reflects something that is real. In our dataset, the *squared* true score weight could not add much to the prediction because – due to our method of obtaining true scores – all true score weights were positive. However, in other datasets, true score weights might well be negative, depending on which variables are selected as the true scores. Thus, our prediction that the squared true score weight is the relevant predictor applies to the more general case. As predicted, we also found a significant standardized Beta (.06) for the interaction between the squared attitude weight a_i^2 and the attitude correlation r_{AA}. Therefore, inter-rater agreement was in fact higher the more the item assessed evaluative attitudes *and* the perceivers' evaluative attitudes regarding the individual targets converged. In addition, the squared attitude weight also made a strong independent prediction by itself (.08), which was not predicted by our model. We think this effect may be explained in terms of the average attitude correlation (r = 0.12) in the present data set (i.e., inter-rater agreement was higher for more evaluative items, because the perceivers tended to agree with one another in regard to how much they liked the individual targets).

3. Discussion

We presented an algebraic model that incorporates the major insights from the research literature on "substance" and

"evaluation" in person judgment. Our main goal in doing so was to demonstrate how various strands of research on these matters may basically be traced back to the same simple idea. By explicating and formalizing that idea for the first time, we hope to improve conceptual and computational clarity in this field, and to enable a better cumulative integration of research findings in the future. Replicating previous research (e.g., Leising et al., 2014), we showed that the rated social desirability of an item is largely the same as the extent to which responses to that item are influenced by the perceivers' evaluative attitudes toward the targets (cf. McCrae & Costa, 1983).

3.1. Internal consistency

One of the major conclusions that may be drawn from our analyses concerns internal consistency: When a perceiver judges targets on two different items, the correlation between the items will be partly predictable from the interaction of the items' evaluative tones alone. That is because an item's evaluative tone directly mirrors the extent to which it will reflect the perceiver's evaluative attitudes toward the targets, and the perceiver's evaluative attitudes are the same in both assessments (assuming temporal stability). This finding is highly relevant to the controversial debate over the so-called "general factor of personality" (GFP) (Musek, 2007), which basically focuses on the question of whether the first (second-order) factor in factor analyses of Big Five items or scales reflects substance (i.e., there is a very broad personality factor that distinguishes people with more desirable personality features from those with less desirable personality features), or style (i.e., the factor is accounted for by the perceivers' more positive or negative attitudes toward the targets). Numerous studies have shown that a relatively strong common factor exists at the top of the hierarchy of personality factors (for an overview, see Just, 2011). Often, the number of studies, the number of measures, or the number of subjects used in these studies is cited as evidence in favor of this factor's "substantial" nature. However, our model highlights the possibility that – despite high numbers of studies, measures, and subjects – this factor may still be at least partly accounted for by the respective perceivers' evaluative attitudes toward the targets. As long as all data in a study come from the same source (most studies used self-report measures exclusively), substance and evaluation may not be distinguished from one another. Several studies show that, once person descriptions are obtained across different sources (e.g., self versus peer-ratings), the general factor of personality is at least substantially weakened (e.g., Danay & Ziegler, 2011; Riemann & Kandler, 2010). This, however, does not imply that the attitude variance in personality ratings is meaningless or irrelevant. For example, a tendency to view oneself more positively or negatively may be highly consequential for an individual's life, as it may determine which challenges a person is willing to take up. More research is needed to clarify just *how much* of the general factor of personality is rooted in attitude variance, what the *remaining* variance reflects, and how the evaluative and the substantial components of the general factor relate to each other.

It should be noted that our model is not only applicable to data structures like the one that we used for illustrative purposes in the present paper (one perceiver p describes all targets t), but also to the more general case where p is a *group* of perceivers (i.e., each target is described by a different perceiver, but by the same perceiver in both assessments i and j). The only difference would be that in the first case perceiver effects in evaluative attitudes (cf. Srivastava, Guglielmo, & Beer, 2010; Wood, Harms, & Vazire, 2010) would not play a role (i.e., they would affect a given perceiver's judgments of all targets equally and thus not affect the correlation r_{YY}), whereas in the second case they would (i.e., they would *increase* the internal consistency correlation r_{YY}). Relationship effects in evaluative attitudes (i.e., a given perceiver favoring some targets over other targets), however, will always play a role, as long as the items are evaluative: The stronger these relationship effects, the more internal consistency will increase.

It is important to consider the consequences of these effects in regard to scale score computation. As is widely known, aggregating across items tends to improve the reliability of measures, because the covariances between the systematic components of the individual items but not the covariances involving unsystematic error (which are zero) will contribute to the overall scale score variance. The question is whether an overall scale score reliably reflects substance or evaluation, because *both* of these are sources of systematic variance. The answer to that question depends on the extent to which the individual items reflect each component. We will discuss the two most extreme cases conceivable: In the first case, all individual items measure substance ($t_i <> 0$) but no evaluation ($a_i = 0$), and the substance they measure is the same ($r_{TT} = 1$). Under such circumstances, aggregating across items will yield an increasingly reliable score that only reflects true score variance and error variance. In the second case, all individual items measure the perceivers' evaluative attitudes ($a_i <> 0$), but no substance ($t_i = 0$). In this case, aggregating across items will also yield an increasingly reliable score, but the score will only reflect attitude variance and error variance. To conclude: A scale may have high internal consistency even if its items measure no real quality of targets at all, but only the perceivers' evaluative attitudes.

3.2. Inter-rater agreement

Another major conclusion that may be drawn from the present analyses is that inter-rater agreement may partly be predicted from the extent to which evaluative attitudes toward the targets are shared between perceivers. Thus, our model suggests a major refinement as compared to previous conceptualizations in which it was assumed that more evaluative items would always yield *lower* inter-rater agreement (e.g., John & Robins, 1993). According to our model, evaluative items will yield lower inter-rater agreement if the perceivers' attitudes regarding the targets are unrelated, because such attitudes will operate as error variance and dilute the overall correlation. If the perceivers' attitudes are negatively correlated (e.g., Pam prefers male targets and Peter prefers female targets), the overall correlation between two measures might even turn out negative. Positively correlated perceiver attitudes on the other hand might result in *higher* inter-rater agreement. All of this, however, will only be the case to the extent that the item actually reflects the perceivers' evaluative attitudes (a_i^2).

This insight has important implications in regard to the common (mis-)interpretation of inter-rater agreement in terms of validity. If different ratings of the same targets agree with one another, this is often interpreted as evidence in favor of their validity. However, whereas it is true that judgments will tend to agree better with one another the more each judgment is valid, the reverse is not true (for an empirical demonstration, see Borkenau, Leising, & Fritz, 2014). In fact, high inter-rater agreement may be completely based on evaluative attitudes that are shared between perceivers. For example, if Pam and Peter are completely unqualified to infer the intelligence of targets, but both hate men, their judgments of the intelligence of a group of male and female targets might still correlate perfectly with one another, given that "intelligent" is one of the most evaluative terms in the natural person-descriptive lexicon (Anderson, 1968; Leising et al., 2012).

3.3. Validity

Even though we did not directly address issues of validity in the present paper, our model allows for making predictions in that

regard, too. Let us consider the case in which ratings of targets on a personality item (e.g., "intelligent") are used to predict some criterion variable (e.g., intelligence test results) that is supposed to reflect the targets' actual standings on the respective trait. That criterion variable Y_{ct} may also be decomposed in terms of our model. Note that there is no independent perceiver here (i.e., item and "perceiver" are confounded), which is why we only use two indices, one for the targets (t) and one for the criterion variable (c). Ideally, the criterion variable would only reflect true score variance, and probably some measurement error. As a result, validity would *decrease* the more the predictor variable contains evaluation variance (i.e., the more the perceiver's attitudes toward the targets differ, and the more the item reflects those attitudes), because this evaluation variance would not be shared between measures, thus attenuating the overall correlation between them.

It is possible, however, to think of the criterion variable as also containing an evaluation component, which allows us to address issues of (e.g.) test fairness: Just like individual perceivers, a test may have different "attitudes" A_{ct} regarding certain types of targets (e.g., a tendency to advantage or disadvantage men). If we conducted a study of judgmental accuracy, using a sample of perceivers whose evaluative attitudes (A_{pt}) correlate positively with the bias (A_{ct}) of the test that is used as the criterion variable, then the predictor–criterion correlation would be inflated, suggesting better accuracy but actually reflecting only the influence of shared bias.

In her Self-Other-Knowledge Asymmetry (SOKA) model, Vazire (2010) hypothesized that "others know more than the self about highly evaluative traits and this asymmetry is reduced or reversed for evaluatively neutral traits" (p. 285). For the case where an unbiased criterion variable is available, our own model allows us to refine that formulation as follows: For evaluatively neutral traits, the validity of self- and other-judgments should be relatively similar (because individual differences in perceiver attitudes should not affect either of these judgments). For evaluative traits, the data source (i.e., self or other) that has the greater attitude variance should be less valid, because this variance would operate as error variance, attenuating the predictor–criterion correlation. For example, if the others (e.g., unacquainted observers) have no reason to prefer some targets over other targets, the attitude variance in the self-ratings (i.e., self-esteem) is likely to be larger (Furr & Funder, 1998; Judge et al., 2002), and thus self-ratings may be less accurate than other-ratings. However, if the sample of others contains individuals whose attitudes regarding their respective targets differ considerably (e.g., spouses and ex-spouses), the attitude variance in the other-ratings may be larger than in the self-ratings, and thus other-ratings may be less accurate than self-ratings.

3.4. Possible extensions of the model

The model we formulated and the applications of the model that we discussed so far were intentionally kept simple in order to make the presentation as easy to follow as possible. However, the model may easily be modified to incorporate additional factors featured in the person perception literature. We will now briefly address some of these.

3.4.1. Decomposing the attitude weight a_i

The extent to which an item reflects the perceivers' attitudes a_i may be decomposed into a normative and an individual component (cf. Borkenau, Zaltauskas, & Leising, 2009). The *normative* component would reflect the extent to which an *average* person would use the item to express a positive or negative evaluation of a target. In contrast, the *individual* component would reflect the extent to which a *specific* perceiver would use the item to express a positive or negative evaluation of a target. By making this distinction, it becomes possible that perceivers who have the same evaluative attitudes toward the same targets may still arrive at different judgments because they differ in how positive or negative a connotation they think an item has. For example, if Priscilla and Phoebe are highly – and equally – fond of Todd, but Priscilla thinks that the word "smartass" is very derogative, whereas Phoebe thinks that the word is only somewhat derogative, then we would expect Priscilla to judge Todd as being *less* of a "smartass", as compared to how Todd is judged by Phoebe. More generally speaking, we would expect the correlation between judgments to be partly predictable from how much a_i is shared between perceivers. In a yet unpublished study of ours, we let a convenience sample of 34 research participants (age: $M = 29.2$, $SD = 9.7$; 16 female, 13 male, 5 failed to report sex) judge the social desirability of 118 questionnaire items. The rating scale ranged from 1 ("very negative") to 10 ("very positive"). As is common in this type of research, the internal consistency of the average desirability judgment was close to perfect, $ICC(3,34) = .99$. However, agreement between *individual* raters was considerably lower, $ICC(3,1) = .73$. Moreover, men ($M = 5.36$, $SD = 0.28$) judged the items more positively than did women ($M = 5.02$, $SD = 0.31$), $t(27) = 3.05$, $p = .005$. Although this latter finding may not be generalizable due to the small sample size and the sampling procedure, we report it here because it suggests that searching for systematic *group differences* in how items are evaluated may be a worthwhile endeavor.

3.4.2. Decomposing A_{pt}

In the present paper, we operationalized a perceiver's evaluative Attitude in terms of how much the perceiver said he or she *liked* a target. However, not all of the perceivers' evaluative attitudes have to be consciously accessible. It would be conceivable to further decompose A_{pt} into an explicit and an implicit component, or into a voluntary (controlled) and an involuntary (automatic) component. For example, Sackeim and Gur (1979) introduced the important distinction between overly positive images of the self that the self believes in (self-deception) and overly positive images of the self that the self knows are inaccurate (other-deception or impression management; cf. Paulhus, 1984). One could conceive of A_{pt} as being a weighted sum of these two influences. The correlation between two measurements would then at least partly depend on the extent to which they reflect each of these influences, and also on the covariations of these influences (e.g., self-other agreement would increase if the targets' unconscious self-deception is "bought" by others).

3.5. Limitations

The present paper clearly has its limitations. First, we used ratings that were averaged across liking levels as our measure of the targets' "true scores". Although this approach is conceptually sound and largely yielded the expected results, it would be preferable to use measures of true scores that are obtained independent of Y_{pti} (e.g., results of cognitive ability tests). Only under such more ideal conditions could we expect the three-way interaction between t_i, t_j, and r_{TT} to influence internal consistency, as predicted by the model. Second, the targets in our empirical present study were public figures the perceivers did not have any personal contact with. As most person judgments in everyday life concern personal acquaintances of the perceivers, this design feature may cast some doubt on the generalizability of the results. Future research certainly needs to replicate the main conclusions of the present study using more naturalistic designs. Third, we used an explicit assessment of Liking (How much do you like this person?) as our measure of A_{pt}. In the context of the present study, this seemed justifiable because all participants judged the same set of target persons. However, Liking under less controlled conditions (e.g., where

each perceiver judges a different target) probably reflects the perceivers' idiosyncratic evaluative attitudes, *as well as* the targets' actual personalities (e.g., in terms of "likeability"). In other words, the perceivers' liking may be correlated with T_{ti}. This, however, does not really pose a problem for the model because it is capable of incorporating correlated predictors (cf. West & Kenny, 2011).

On a side note: Due to the close correspondence between the item's evaluative tone and the perceiver's evaluative attitude, the extent to which a perceiver endorses positive and does not endorse negative items in describing a target may be used as an indirect but relatively accurate measure of the perceiver's evaluative attitude or Liking. This is particularly true when aggregating across many items, as is the case when computing profile correlations between item endorsements and item desirabilities for a given perceiver's description of a given target. For example, in Leising et al.'s (2013) study, such profile correlations correlated very strongly ($r(3) = .92$, $p < .05$) with the Liking level (1–5) reported by the respective perceiver (correlation not reported in the original study).

3.6. Conclusion

This paper presented an algebraic formulation of some basic ideas that have permeated the literature on substance and evaluation in person perception for decades. It should be noted that we by no means claim to have come up with these ideas ourselves. Rather, the most important conceptual contributions may clearly be attributed to other authors, such as Edwards (1953), Peabody (1967), McCrae and Costa (1983), John and Robins (1993) and Bäckström et al. (2009) (see Introduction). The merit of the present paper is that it outlines these ideas in a precise mathematical fashion for the first time. By providing this outline, we hope to foster the integration of the various existing strands of research addressing issues of substance and evaluation. Such integration has often been hampered by inconsistent terminology and methodology, resulting in a vast number of different but overlapping concepts and findings that are often hard to comprehensively conciliate. The algebraic model presented here should be able to help streamline the interpretation of the existing body of evidence considerably.

References

Anderson, N. (1968). Likeableness ratings of 555 personality-trait words. *Journal of Personality and Social Psychology, 9*, 272–279. http://dx.doi.org/10.1037/h0025907.

Bäckström, M., Björklund, F., & Larsson, M. R. (2009). Five-factor inventories have a major general factor related to social desirability which can be reduced by framing items neutrally. *Journal of Research in Personality, 43*, 335–344. http://dx.doi.org/10.1016/j.jrp.2008.12.013.

Bono, J. E., & Judge, T. A. (2003). Core self-evaluations: A review of the trait and its role in job satisfaction and job performance. *European Journal of Personality, 17*, 5–18. http://dx.doi.org/10.1002/per.481.

Borkenau, P., Leising, D., & Fritz, U. (2014). Effects of communication between judges on consensus and accuracy in judgments of people's intelligence. *European Journal of Psychological Assessment, 30*, 274–282. http://dx.doi.org/10.1027/1015-5759/a000188.

Borkenau, P., & Ostendorf, F. (1989). Descriptive consistency and social desirability in self-and peer reports. *European Journal of Personality, 3*, 31–45. http://dx.doi.org/10.1002/per.2410030105.

Borkenau, P., & Ostendorf, F. (1998). The Big Five as states: How useful is the five-factor model to describe intraindividual variations over time? *Journal of Research in Personality, 32*, 202–221. http://dx.doi.org/10.1006/jrpe.1997.2206.

Borkenau, P., & Zaltauskas, K. (2009). Effects of self-enhancement on agreement on personality profiles. *European Journal of Personality, 23*, 107–123. http://dx.doi.org/10.1002/per.707.

Borkenau, P., Zaltauskas, K., & Leising, D. (2009). More may be better, but there may be too much. Optimal trait level and self-enhancement bias. *Journal of Personality, 77*, 825–858. http://dx.doi.org/10.1111/j.1467-6494.2009.00566.x.

Danay, E., & Ziegler, M. (2011). Is there really a single factor of personality? A multirater approach to the apex of personality. *Journal of Research in Personality, 45*, 560–567. http://dx.doi.org/10.1016/j.jrp.2011.07.003.

Dumas, J. E., Johnson, M., & Lynch, A. M. (2002). Likableness, familiarity, and frequency of 844 person-descriptive words. *Personality and Individual Differences, 32*, 523–531. http://dx.doi.org/10.1016/S0191-8869(01)00054-X.

Edwards, A. L. (1953). The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. *Journal of Applied Psychology, 37*, 90–93. http://dx.doi.org/10.1037/h0058073.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. London, UK: Chapman & Hall.

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

Furr, R. M., & Funder, D. C. (1998). A multimodal analysis of personal negativity. *Journal of Personality and Social Psychology, 74*, 1580–1591. http://dx.doi.org/10.1037//0022-3514.74.6.1580.

John, O. P., & Robins, R. W. (1993). Determinants of interjudge agreement on personality traits: The Big Five domains, observability, evaluativeness, and the unique perspective of the self. *Journal of Personality, 61*, 521–551. http://dx.doi.org/10.1111/j.1467-6494.1993.tb00781.x.

Judge, T. A., Erez, A., Bono, J. E., & Thoresen, C. J. (2002). Are measures of self-esteem, neuroticism, locus of control, and generalized self-efficacy indicators if a common core construct? *Journal of Personality and Social Psychology, 83*, 693–710. http://dx.doi.org/10.1037//0022-3514.83.3.693.

Just, C. (2011). A review of literature on the general factor of personality. *Personality and Individual Differences, 50*, 765–771. http://dx.doi.org/10.1016/j.paid.2011.01.008.

Kenny, D. A. (2004). PERSON: A general model of interpersonal perception. *Personality and Social Psychology Review, 8*, 265–280. http://dx.doi.org/10.1207/s15327957pspr0803_3.

Kenny, D. A., Albright, L., Malloy, T. E., & Kashy, D. A. (1994). Consensus in interpersonal perception: Acquaintance and the Big Five. *Psychological Bulletin, 116*, 245–258. http://dx.doi.org/10.1037/0033-2909.116.2.245.

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

Kenny, D. A., & West, T. V. (2010). Similarity and agreement in self-and other perception: A meta-analysis. *Personality and Social Psychology Review, 14*, 196–213. http://dx.doi.org/10.1177/1088868309353414.

Leising, D., & Borkenau, P. (2011). Person perception, dispositional inferences and social judgment. In *Handbook of interpersonal psychology: Theory, research, assessment, and therapeutic interventions* (pp. 157–170). Hoboken, NJ: John Wiley & Sons.

Leising, D., Borkenau, P., Zimmermann, J., Roski, C., Leonhardt, A., & Schütz, A. (2013). Positive self-regard and claim to leadership: Two fundamental forms of self-evaluation. *European Journal of Personality, 27*, 565–579. http://dx.doi.org/10.1002/per.1924.

Leising, D., Erbs, J., & Fritz, U. (2010). The letter of recommendation effect in informant ratings of personality. *Journal of Personality and Social Psychology, 98*, 668–682. http://dx.doi.org/10.1037/a0018771.

Leising, D., Gallrein, A. M. B., & Dufner, M. (2014). Judging the behavior of people we know: Objective assessment, confirmation of preexisting views, or both? *Personality and Social Psychology Bulletin, 40*, 153–163. http://dx.doi.org/10.1177/0146167213507287.

Leising, D., Ostrovski, O., & Borkenau, P. (2012). Vocabulary for describing disliked persons is more differentiated than vocabulary for describing liked persons. *Journal of Research in Personality, 46*, 393–396. http://dx.doi.org/10.1016/j.jrp.2012.03.006.

Leising, D., Ostrovski, O., & Zimmermann, J. (2013). "Are we talking about the same person here?" Interrater agreement in judgments of personality varies dramatically with how much the perceivers like the targets. *Social Psychological and Personality Science, 4*, 468–474. http://dx.doi.org/10.1177/1948550612462414.

Locke, K. D., & Christensen, L. (2007). Re-construing the relational-interdependent self-construal and its relationship with self-consistency. *Journal of Research in Personality, 41*, 389–402. http://dx.doi.org/10.1016/j.jrp.2006.04.005.

Locke, K. D., Craig, T., Baik, K.-D., & Gohil, K. (2012). Binds and bounds of communion: Effects of interpersonal values on assumed similarity of self and others. *Journal of Personality and Social Psychology, 103*, 879–897. http://dx.doi.org/10.1037/a0029422.

Locke, K. D., Zheng, D., & Smith, J. (2014). Establishing commonality versus affirming distinctiveness: Patterns of personality judgments in China and the United States. *Social Psychological and Personality Science, 5*, 389–397. http://dx.doi.org/10.1177/1948550613506718.

McCrae, R. R., & Costa, P. T. (1983). Social desirability scales: More substance than style. *Journal of Consulting and Clinical Psychology, 51*, 882–888. http://dx.doi.org/10.1037/0022-006X.51.6.882.

Musek, J. (2007). A general factor of personality: Evidence for the Big One in the five factor model. *Journal of Research in Personality, 41*, 1213–1233. http://dx.doi.org/10.1016/j.jrp.2007.02.003.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46*, 598–609. http://dx.doi.org/10.1037/0022-3514.46.3.598.

Peabody, D. (1967). Trait inferences: Evaluative and descriptive aspects. *Journal of Personality and Social Psychology, 7*, 1–18. http://dx.doi.org/10.1037/h0025230.

Riemann, R., & Kandler, C. (2010). Construct validation using multitrait-multimethod-twin data: The case of a General Factor of Personality. *European Journal of Personality, 24*, 258–277. http://dx.doi.org/10.1002/per.760.

Sackeim, H. A., & Gur, R. C. (1979). Self-deception, other-deception, and self-reported psychopathology. *Journal of Consulting and Clinical Psychology, 47*, 213–215. http://dx.doi.org/10.1037/0022-006X.47.1.213.

Saucier, G. (1994). Separating description and evaluation in the structure of personality attributes. *Journal of Personality and Social Psychology, 66*, 141–154. http://dx.doi.org/10.1037//0022-3514.66.1.141.

Srivastava, S., Guglielmo, S., & Beer, J. S. (2010). Perceiving others' personalities: Examining the dimensionality, assumed similarity to the self, and stability of perceiver effects. *Journal of Personality and Social Psychology, 98*, 520–534. http://dx.doi.org/10.1037/a0017057.

Vazire, S. (2010). Who knows what about a person? The self-other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology, 98*, 281–300. http://dx.doi.org/10.1037/a0017908.

West, T. V., & Kenny, D. A. (2011). The truth and bias model of judgment. *Psychological Review, 118*, 357–378. http://dx.doi.org/10.1037/a0022936.

Wood, D., Harms, P., & Vazire, S. (2010). Perceiver effects as projective tests: What your perceptions of others say about you. *Journal of Personality and Social Psychology, 99*, 174–190. http://dx.doi.org/10.1037/a0019390.