European Journal of Personality, *Eur. J. Pers.* **30**: 618–636 (2016)
Published online in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/per.2083

Using Personality Item Characteristics to Predict Single-Item Internal Reliability, Retest Reliability, and Self–Other Agreement

REINOUT E. DE VRIES[1,2]*, ANU REALO[3,4] and JÜRI ALLIK[4,5]

[1]*Vrije Universiteit Amsterdam, The Netherlands*
[2]*University of Twente, The Netherlands*
[3]*University of Warwick, UK*
[4]*University of Tartu, Estonia*
[5]*Estonian Academy of Sciences, Estonia*

Abstract: The use of reliability estimates is increasingly scrutinized as scholars become more aware that test–retest stability and self–other agreement provide a better approximation of the theoretical and practical usefulness of an instrument than its internal reliability. In this study, we investigate item characteristics that potentially impact single-item internal reliability, retest reliability, and self–other agreement. Across two large samples (N= 6690 and N= 4396), two countries (Estonia and The Netherlands), and two personality inventories (the NEO PI-3 and the HEXACO-PI-R), results show that (i) item variance is a strong predictor of self–other agreement and retest reliability but not of single-item internal reliability; (ii) item variance mediates the relations between evaluativeness and self–other agreement; and (iii) self–other agreement is predicted by observability and item domain. On the whole, weak relations between item length, negations, and item position (indicating effects of questionnaire length) on the one hand, and single-item internal reliability, retest reliability, and self–other agreement on the other, were observed. In order to increase the predictive validity of personality scales, our findings suggest that during the construction of questionnaire items, researchers are advised to pay close attention especially to item variance, but also to evaluativeness and observability. Copyright © 2016 European Association of Personality Psychology

Key words: reliability; self–other agreement; evaluativeness; five-factor model; HEXACO

Recent articles by McCrae and colleagues (McCrae, 2015; McCrae, Kurtz, Yamagata, & Terracciano, 2011) have highlighted an 'inconvenient psychometric truth' in personality assessment, namely that internal reliability (often referred to as 'internal consistency')[1] is not an adequate predictor of criterion validity, whereas retest reliability and self–other agreement are. Although it is becoming increasingly apparent that information about retest reliability and self–other agreement is more useful for scale development and validation purposes than information about internal reliability, it is less clear what properties items should have to ensure sufficient high levels of retest reliability and self–other agreement. To elucidate this question, it is necessary to focus on the predictors of internal reliability, retest reliability, and self–other agreement instead of on their consequences. Thus, in contrast to McCrae et al. (2011), who focused on the *effects* of internal reliability and retest reliability on the predictive validity of personality traits at the *facet and domain* level, in this research, we will focus instead on the *predictors* of internal reliability, retest reliability, and self–other agreement at the *item level*. That is, we will investigate if—and to what extent—different item characteristics predict internal reliability (estimated by single-item internal reliability (Wanous and Reichers (1996)), retest reliability (only in Study 2), and self–other agreement.

An inconvenient truth

The finding by McCrae et al. (2011; 2015) that self–other agreement and retest reliability are better predictors of criterion validity than internal reliability is inconvenient because, according to a ruling dogma among practitioners and researchers, internal reliability is a necessary, albeit insufficient, condition for ensuring criterion-related validity. The first part of this dogma—that internal reliability is a necessary condition for validity—has become so deeply entrenched in our field that instruments with insufficient internal reliability levels (according to the dogma, usually lower than .70) are regarded with suspicion, and that articles which report such instruments may be rejected for this reason alone. That is, when a dearth of validity information is present, researchers and reviewers may commit the Type I fallacy of rejecting a potentially valid instrument because

*Correspondence to: Reinout E. de Vries, Vrije Universiteit Amsterdam, Department of Experimental and Applied Psychology, Van der Boechorststraat 1, 1081 BT, Amsterdam, The Netherlands. E-mail: re.de.vries@vu.nl
Contract/grant sponsor: Estonian Ministry of Education and Science; contract/grant number: IUT2-13.
[1]We use the term 'internal reliability' instead of the more commonly used term 'internal consistency reliability' because of research that has shown that the latter is actually a misnomer. Measures can actually have low levels of internal consistency (i.e. because of multidimensionality) and still have high levels of internal reliability, such as shown, for instance, by high levels of Cronbach alpha (Sijtsma, 2015).

Copyright © 2016 European Association of Personality Psychology
Received 25 January 2016
Revised 17 October 2016, Accepted 19 October 2016

of its low internal reliability. At the same time, in the absence of validity information or when validities are tainted by common-method biases, researchers and reviewers may be tempted to commit the Type II fallacy of accepting a reliable but invalid instrument. In leadership research, for instance, there has been a longstanding critique of instruments which contain items that are highly evaluative—ensuring high levels of internal reliability—but on which relatively low levels of self–other and other-other agreement are observed (e.g. De Vries, 2012; Ostroff, Atwater, & Feinberg, 2004; Warr & Bourne, 1999), and which, as a consequence, may have lower levels of predictive validity (e.g. Atwater & Yammarino, 1992; McCrae et al., 2011; Whittington, Coker, Goodwin, Ickes, & Murray, 2009).

In personality research, this debate has found its zenith in the discussion on the usefulness of short and mostly unreliable personality scales, with research showing that personality scales can still attain sufficient levels of validity even in the presence of low levels of internal reliability (Burisch, 1997; De Vries, 2013; Thalmayer, Saucier, & Eigenhuis, 2011). In fact, it has become increasingly recognized in personality research that the true necessary, but insufficient, condition for predictive criterion-related validity is retest reliability (McCrae et al., 2011) or interrater (e.g. self–other or other-other) agreement.[2] At the same time, the use of internal reliability estimates has become criticized because high levels of internal reliability may be caused by the presence of 'bloated specifics' (Cattell, 1973), transient errors (Becker, 2000; Chmielewski & Watson, 2009; Schmidt, Le, & Ilies, 2003; Thorndike, 1951), or method variance (McCrae, 2015). That is, whereas method variance, associated with response styles or socially desirable responding, may impact reliability positively, it may also be associated with lower levels of predictive validity.

Single-item internal reliability, retest reliability, and self–other agreement

Arguably, based on the above, the adequacy of a personality item should be based more on its (single-item) self–other agreement and retest reliability than on its single-item internal reliability. Consequently, the main question for this research is: What item characteristics affect single-item internal reliability, retest reliability, and self–other agreement? To answer this question, first of all, we have to specify what we mean with single-item internal reliability, (single-item) retest reliability, and (single-item) self–other agreement and, second, we need to discuss how these three components are related to each other. To start with the meaning of these three components, of these three, retest reliability and self–other agreement are most easily conceptualized. They are conceptualized respectively using the correlation between an item measured at T1 and the same item measured at T2 (retest reliability) and the correlation between an item's self- and observer-ratings (self–other agreement; note, we will drop the 'single-item' in front of retest reliability and self–other agreement from here on but we retain it when referring to single-item internal reliability). Single-item internal reliability, however, can be conceptualized in two different ways. The first conceptualization is based on the single-item (internal) reliability formula proposed by Wanous and Reichers (1996), that is, $r_{xx} = r_{xy}^2/r_{yy}$, in which r_{xx} is the single-item internal reliability, r_{xy} is the item-rest correlation between the item and its scale, and r_{yy} is the internal reliability of the scale without the target item (i.e. reliability if the target item is deleted). The second conceptualization is based on the communality or the squared factor loading of a single item when combined with other items that represent the same construct in principal axis factoring. When principal axis factoring is performed and only one factor is extracted, the squared loading—which is equal to the communality—of a single item reflects the amount of shared variance of the item with the underlying factor or its proportion of true variance, which can thus be equated to its reliability (Denissen, Geenen, Selfhout, & Van Aken, 2008; Spörrle & Bekk, 2014; Wanous & Hudy, 2001). Single-item internal reliabilities are most often used in surveys that can only take a very short time because people are unwilling or unable to answer lengthy questionnaires (e.g. on the street or in the hospital) or when the survey involves a very large number of scales, such as in large-scale social surveys. They are less often used in personality research, because, except maybe for Extraversion, most single-item measures of personality are characterized by low levels of single-item internal reliabilities (Spörrle & Bekk, 2014). In this study, we combine both conceptualizations of single-item internal reliability into one single-item internal reliability estimate.

Second, the three components—single-item internal reliability, retest reliability, and self–other agreement—do have some overlap. To deduce the amount of overlap between the three components, McCrae (2015) posited that the variance of an item can be partitioned into trait-related (systematic) variance (T), method variance (M), specific (item-related) variance (s), and error variance (ε). Trait variance is that part of the variance that is because of systematic and 'true' individual differences in the (personality) characteristic being assessed. Method variance is because of individual differences in response biases. That is, some respondents may tend to exaggerate their standing on desirable traits and downplay their standing on undesirable traits, whereas others do not or even react oppositely. Specific variance is that part of the variance of an item that is not shared by the other items in the scale. That is, it refers to 'true' individual differences in item-specific variance. And finally, error variance is random variance or 'noise' and refers to that part of the variance that is unsystematic and (by definition) unrelated to the other sources of variance.

According to McCrae (2015), internal reliability (α) is a function of both the amount of trait and method variance, that is, $\alpha = (T + M)/(T + M + s + \varepsilon)$; retest reliability ($r_{tt}$) is a

[2]First, note that we define validity in terms of a predictor-criterion relation, in which the predictor is a personality trait and the criterion is a proximate or ultimate outcome, such as school success, therapy outcome, relationship satisfaction, or work performance. Second, note that interrater agreement may be less feasible for some other areas of psychological research, that is, when investigating internal, short-term states or attitudes.

Copyright © 2016 European Association of Personality Psychology

DOI: 10.1002/per

function of the amount of trait, method, and specific variance, that is, $r_{tt} = (T + M + s)/(T + M + s + \varepsilon)$[3]; and self–other agreement (r_{so}) is a function of the amount of trait and specific variance, that is, $r_{so} = (T + s)/(T + M + s + \varepsilon)$. Thus, retest reliability overlaps with both internal reliability and self–other agreement because it shares with the former trait variance (T) and method variance (M) and with the latter trait variance (T) and specific variance (s). But, although both internal reliability and self–other agreement depend on the amount of trait variance, the main difference between internal reliability and self–other agreement is that besides trait variance (T), internal reliability is also a function of method variance (M) whereas self–other agreement is also a function of specific variance (s).

In the Revised NEO Personality Inventory (NEO PI-R; Costa & McCrae, 1992), the mean prevalence of these four variance components at the item level was estimated to be 12% (trait variance), 13% (method variance), 24% (specific variance), and 51% (error variance) (McCrae, 2015; pp. 105–106). That is, specific variance contributed almost a quarter of the total variance whereas trait and method variance each contributed one-eight approximately. According to McCrae (2015), items with a large amount of specific variance relative to other sources of variance have higher levels of self–other agreement and retest reliability, but lower levels of internal reliability. In contrast, items with a large amount of method variance relative to the other sources of variance have higher levels of internal reliability and retest reliability, but lower levels of self–other agreement. That is, large amounts of method variance are likely to result in an unjustified raise in (internal and retest) reliability but may, at the same time and to the extent that it replaces systematic and specific variance, result in an attenuation of self–other agreement.

Domains of item characteristics

What kind of item characteristics determines single-item internal reliability, retest reliability, and self–other agreement? According to the Realistic Accuracy Model of Funder (1995), to be able to accurately assess personality traits, a target needs to exhibit behaviours that are relevant to a given trait (relevance), which are 'visibly' expressed and thus become available for observers (availability). In turn, judges (either the target him-/herself or an external observer) need to be able to detect the relevant and available behaviours (detection), which in turn need to be correctly utilized as indicative of a trait (utilization). Although Funder (1995) conceptualized his model as a *person characteristics* model, in which (information about traits of) targets and judges play a central role, the four process variables of his model can be readily adapted to an *item characteristics* model. That is, item characteristics may determine whether the behaviour described in an item is relevant for a trait, whether the behaviour described is usually available to judges, whether judges will be able to detect the behaviour described in the item, and whether judges will correctly utilize the information.

Here, we make a distinction between seven item characteristics—that is, item variance, evaluativeness, observability, item domain, item position, item length, and negation—that may impact item relevance, availability, detection, and utilization. The two item characteristics associated with item *relevance* are item variance and evaluativeness. According to Funder (1995, p. 658), 'in some context or contexts, a trait produces a behavioural effect. The resulting behaviour is then relevant to that trait.' An item can be seen as a context which invites a behavioural response based on somebody's trait level. The more an item is associated with instances in which a trait gets activated (Tett & Burnett, 2003), the more varied the response, the higher the variance, and the more relevant the item is for the trait in question. Thus, item variance is an important manifestation of item relevance. Evaluativeness may affect item relevance, because high negative or positive evaluativeness (or: low or high social desirability) of the behaviour in an item might redirect a judge's attention away from the trait in the question, decreasing the relevance of the item for the trait that is measured.

The two item characteristics associated with *availability* are observability of the behaviour and item domain. Observability affects availability because items that contain statements on behaviours that are more readily observable are by definition more 'available' to an external observer than items that contain statements on behaviours that are not easily observable. Item domain is associated with availability because some personality domains, such as Extraversion, are more readily expressed in behaviours than other personality domains (Funder & Dobroth, 1987; Watson, Hubbard, & Wiese, 2000) and thus it may be easier to write items with observable content for these domains than for other domains.

The item characteristic associated with *detection* is item position in the questionnaire. Item position may affect detection, because respondents may get tired or bored at the end of a long questionnaire, and thus items positioned at the end of a such a questionnaire may be processed suboptimally, interfering with the detection of the behaviour involved in the item (Galesic & Bosnjak, 2009; Herzog & Bachman, 1981; Kraut, Wolfson, & Rothenberg, 1975). And finally, the two item characteristics associated with (correct) *utilization* are item length and negation. Item length and negation are associated with utilization because longer items and items that contain negations may be more complex and confusing, thus introducing judgement errors (Marsh, 1986; Saucier & Golberg, 2002). Thus, whereas detection is associated with item position in a long questionnaire—a characteristic that can be remedied to some extent by randomizing items in such a questionnaire—utilization is associated with the style in which an item is written.

The item characteristics model described here is not a process model such as the RAM (Funder, 1995). Each of the elements described here is likely to have an independent

[3]Whereas trait and specific variance are deemed to be stable, method variance may consist of a stable part (i.e. a similar way of responding to—for instance—socially desirable items across occasions) and an unstable part (i.e. transient errors or differential responding across occasions). Whereas both forms of method variance will result in an increase in internal reliability, only stable method variance will increase retest reliability.

Copyright © 2016 European Association of Personality Psychology

DOI: 10.1002/per

effect on single-item internal reliability, retest reliability, and self–other agreement, although we will argue that evaluativeness will impact these three reliability estimates at least partly through item variance. Below, we will discuss each of the seven item characteristics, item variance, evaluativeness, observability, item domain, item position, item length, and negation, in turn.

Item relevance: Item variance

First of all, we will investigate the effects of item variance on single-item internal reliability, retest reliability, and self–other agreement. Without any item variance, no relevant behaviour is revealed, and, consequently, there is zero single-item internal reliability, retest reliability, and self–other agreement. Thus, higher item variance should be positively related to single-item internal reliability, retest reliability, and self–other agreement. Empirical evidence has shown that facet-level variance is a strong predictor of self–other agreement (Allik et al., 2010). The question is, however, whether item-level variance contributes equally to single-item internal reliability, retest reliability, and self–other agreement. When considering the findings of McCrae (2015) on the mean prevalence of the trait, method, and specific variance components described above, item variance should be least strongly related to single-item internal reliability (which accounted for (12% + 13%=) 25% of the item variance), somewhat more strongly to self–other agreement (which accounted for (12% + 24%=) 36% of the item variance), and most strongly to retest reliability (which accounted for (12% + 13% + 24%=) 49% of the item variance), but this is only true when the relative amounts of these variance proportions are the same for all items. That is, there may actually be more variance across items in the amount of trait variance and specific variance than there is variance across items in the amount of method and error variance. If this is the case, it may be true that item variance is mainly related to self–other agreement, and to a lesser extent to retest reliability and single-item internal reliability.

Item relevance: Evaluativeness (or: Social desirability)

In personality research, items that invite evaluative or socially desirable responding are usually treated with suspicion because such items are associated with higher levels of method variance (Ashton, De Vries, & Lee, in press; Bäckström, Björklund, & Larsson, 2009; John & Robins, 1993). Method variance may consist of variance associated with response styles (e.g. extreme responding, midpoint responding, acquiescent responding; Zettler, Lang, Hülsheger, & Hilbig, 2016) and variance associated with socially desirable responding. Of these two, item characteristics (such as the evaluativeness of an item) are by definition more likely to invite differences in socially desirable responding. That is, when answering a personality item, *a person's socially desirable response* is determined by a confluence of the item content (i.e. whether agreeing or disagreeing with an item is socially desirable or not) and the individual characteristics of the person answering the item. Answers to items with neutral socially desirable content should, by definition, be determined by someone's actual traits (although admittedly, people may differ in the extent to which they think an item is neutral with respect to social desirability, see Wood & Wortman, 2012). Answers to items with non-neutral—that is, socially desirable or undesirable content—are also determined by someone's actual traits, but also by the extent to which a person tends to respond in a socially desirable manner. Consequently, items will differ in the extent to which they invite socially desirable responding. In contrast, *a person's response style* should—by definition—influence *all* personality items equally. Although items with completely different 'content domains' (e.g. items that reflect someone's knowledge versus items that reflect someone's personality) may invite different response styles from the same person, items of the same content domain (in this case, personality) are unlikely to differ in the extent to which they invite differences in responding to different items. That is, response styles, such as midpoint responding, extreme responding, and acquiescence, constitute a 'domain-specific' method factor rather than an 'item-specific' method factor.

Furthermore, extremely desirable or undesirable items (i.e. items high on evaluativeness) are more likely to be associated with higher (or respectively lower) means and, consequently, because of restriction of range, lower item variances. That is, when mean evaluativeness of an item is increased, trait variance is likely to decrease. Research has shown that adjectival items are seldom evaluatively neutral and, consequently, the use of adjectival items has strong effects on self–other agreement, with evaluatively neutral items inviting higher levels of self–other agreement when compared to evaluatively positive or negative items (John & Robins, 1993). The use of questionnaire items is recommended by virtue of the fact that modifiers can be used to change the evaluativeness of statements. But even in statements instead of single-word items, evaluativeness is likely to affect self–other agreement. Research shows, for instance, that questionnaires that invite socially desirable responding are more likely to result in factor scales that are more strongly interrelated, whereas factor scales consisting of items with reduced social desirability are much less strongly interrelated (Bäckström et al., 2009). Psychopathology questionnaires are especially susceptible to this effect, with research showing that self–other agreement—but not reliability—is compromised because of the use in psychopathology questionnaires of highly (negatively valenced) evaluative items (Ashton et al., in press).[4] Thus, deviations from neutral evaluativeness may cause a reduction of item variance, which in turn may lead to lower levels of self–other agreement. That is, item variance is expected to mediate the negative relation between evaluativeness and self–other agreement.

Availability: Observability and item domain

It has long been recognized that observability (or: visibility) of traits may be an important precursor for self–other

[4]See Table 1 of Ashton et al. (in press). In fact, whereas the mean facet-level self–other agreement was lower (e.g. .44 in the Personality Inventory for the DSM-5 (PID-5) versus .52 in the HEXACO-PI-R), the mean facet-level reliability of the PID-5 was higher than that of the HEXACO-PI-R (e.g. .80 versus .65 in self-ratings and .85 versus .68 in observer ratings).

Copyright © 2016 European Association of Personality Psychology

DOI: 10.1002/per

agreement (Funder & Dobroth, 1987; Watson et al., 2000). Basically, observability ensures that the behaviour referred to in the item is available to the rater, which in turn may have a positive effect on the amount of trait and specific variance. Especially Extraversion has been often singled out as the most observable trait, but research suggests that it may depend on situational affordances which traits are activated and which are not, making some traits in some situations more observable than others (De Vries, Tybur, Pollet, & Van Vugt, 2016; Rauthmann, 2012; Tett & Burnett, 2003). An aspect that may be especially relevant in this regard is whether traits are associated with engagement or with altruism, a distinction that has been proffered in the HEXACO model (Ashton & Lee, 2001, 2007; Ashton, Lee, & De Vries, 2014). Engagement traits refer to Extraversion (social engagement), Conscientiousness (task engagement), and Openness to Experience (idea engagement) and Altruism traits refer to Honesty–Humility (reciprocal altruism), Emotionality (kin altruism), and Agreeableness (reciprocal altruism). In the Big Five or Five-Factor model (e.g. Costa & McCrae, 1992; Goldberg, 1990), the engagement factors are similar to those of the HEXACO model, whereas the altruism factors are constituted by rotational variants of HEXACO Agreeableness and Emotionality, that is, Big Five Agreeableness and Emotional Stability (Ashton et al., 2014). Especially traits that are associated with altruism may be somewhat less observable than traits associated with engagement. That is, expressions of low Honesty–Humility, high Emotionality, and low Agreeableness may be somewhat more uncommon because people are less frequently in situations that allow for the expression of these traits. On the whole, because observability is likely to positively impact the amount of trait and specific item variance relative to the amount of method and error variance, observability should be mainly related to self–other agreement and should have less of an effect on single-item internal reliability.

Detection: Item position

It is a common assumption that questionnaire length has a negative effect on response quality because respondents may tend to become tired or bored with long questionnaires, thus being less able and willing to detect and absorb the behavioural content of items positioned at the end. Items placed at the end of a survey have indeed been found to be associated with somewhat less extreme responses (Kraut et al., 1975), more straightlining (Herzog & Bachman, 1981), and faster response times (Galesic & Bosnjak, 2009), indicative of careless responding and thus a higher level of error variance . In this study, we will investigate whether 'item position' (i.e. which position the item has in the questionnaire; from the first to the last item) is related to reliability and self–other agreement estimates. Because item position may potentially lower response quality and because, consequently, item position may be associated with a higher amount of error variance relative to the other sources of variance, we expect single-item internal reliability, retest reliability, and self–other agreement to be negatively affected.

Utilization: Item length

Another item characteristic that we will investigate is item length. Item length has been often considered an important impeding factor of item comprehension, resulting in less correct utilization of an item because of greater item complexity (Hofstee, 1991; Saucier & Golberg, 2002), which, in turn, may result in more error variance and, consequently, lower single-item internal reliability, retest reliability, and self–other agreement. As a consequence, some authors have proposed that personality items should be as short as possible, consisting solely of a verb and a verb specifier (Hendriks, 1997; Hendriks, Hofstee, & De Raad, 1999). Indeed, when comparing NEO PI-R items that have reduced length with items from the original NEO PI-R, the items with reduced length were found to have higher reliabilities (Mõttus, Pullmann, & Allik, 2006). Item length may thus have an effect on both reliability and self–other agreement. It should be noted that the effects of item length may be somewhat curvilinear. That it, very short items—consisting of only one or two words—are often very abstract because of a lack of specification and contextualization and often have higher levels of evaluativeness (e.g. John & Robins, 1993; Wood & Wortman, 2012). The higher level of abstractness may result in higher levels of error variance, which may result in lower levels of reliability and self–other agreement. However, higher levels of evaluativeness may be associated with more method variance, raising reliability levels. As a case in point, scales based on items consisting of single words were found to have lower levels of self–other agreement than scales based on items consisting of short sentences, despite similar levels of internal reliability (Allik et al., 2010; Watson et al., 2000). However, our research does not contain any items that consist of single words and evaluativeness is likely to be less pronounced in item sentences, thus it is most likely that the relations between item length and single-item internal reliability, retest reliability, and self–other agreement are linear and only negatively affected because of the higher level of error variance in longer items.

Utilization: Negation

A related issue concerns the use of negations in items. The use of negations (e.g. 'not,' 'no,' 'nothing,' 'never,' 'less,' 'dis-,' and 'un-') is often considered 'bad practice,' because negations are more likely to result in incorrect utilization of the item because of lower levels of item comprehension (e.g. Hofstee, 1991; McCrae et al., 2011), which, in turn, may result in more error variance and lower internal reliability, retest reliability, and self–other agreement. The negative effects of negations on average item intercorrelations and coefficient alphas have been found to be especially pronounced among young children (Marsh, 1986). Among adults and in single domains such as self–esteem, negations have been found to be related to response styles, which have been related to substantive trait-like factors (DiStefano & Motl, 2006). However, method effects in self-esteem scales because of negations have been associated with higher rather than lower grades among adults (Greenberger, Chen, Dmitrieva, & Farruggia, 2003), suggesting that some of these

Copyright © 2016 European Association of Personality Psychology

DOI: 10.1002/per

effects may be because of conscious self-representation. Whether such effects occur for adults and in broad personality questionnaires instead of narrow constructs such as self-esteem, however, is an empirical question. Such a question can only be addressed if a questionnaire has a sufficient number of items from different trait domains that contain negations in both directions, that is, negations that are associated with socially desirable traits and negations that are associated with socially undesirable traits. Both the (Estonian version of the) NEO PI-3 (McCrae, Costa, & Martin, 2005) and the (Dutch version of the) HEXACO-PI-R (Ashton & Lee, 2008; De Vries, Ashton, & Lee, 2009; Lee & Ashton, 2004) contain a sufficient number of negations (i.e. 71 negations (29.6% of the 240 items) in the Estonian NEO PI-3 and 67 negations (33.5% of the 200 items) in the Dutch HEXACO-PI-R) in both socially desirable and socially undesirable directions (see the negligible correlations between negations and evaluativeness in Study 1 (Table 1) and Study 2 (Table 3)). Consequently, these two questionnaires offer the possibility to inspect the influence of the use of negations on single-item internal reliability, retest reliability (but not in Study 1; see below), and self–other agreement.

THE PRESENT STUDY

In sum, the present study aims to examine if and to what extent the following item characteristics—item variance, evaluativeness, observability, item domain, item position, item length, and negation—predict single-item internal reliability, retest reliability, and self–other agreement. Based on what we argued above, we expect the item relevance and availability variables—item variance, evaluativeness, and observability/item domain—to affect mainly self–other agreement and to a lesser extent—or not at all—single-item internal reliability. Of these, evaluativeness is most likely to affect self–other agreement through item variance. That is, item variance is likely to be reduced—because of restriction of range effects—when items are highly evaluative. Furthermore, we expect the detection and utilization variables—item length, negation, and item position—to affect both reliability and self–other agreement. Items at the end of a questionnaire, longer items, and items with negations are expected to have higher error variance, and as a consequence lower single-item internal reliability, retest reliability, and self–other agreement.

This research is conducted using two large samples containing self- and other-ratings of personality. In the first sample (Study 1), the Estonian version of the NEO PI-3 (McCrae et al., 2005) is used. In the second sample (Study 2), the Dutch version of the HEXACO-PI-R (De Vries et al., 2009; Lee & Ashton, 2004) is used. In both samples, we analysed the data using item characteristics. That is, characteristics such as item variance, evaluativeness, observability, item domain, item position, item length, negation, single-item internal reliability, retest reliability (only in Study 2), and self–other agreement were obtained for each item. Consequently, the analyses are performed on the samples of *items* of the NEO PI-3 and HEXACO-PI-R instead of on the samples of respondents.

STUDY 1

Method

Sample and procedure

The sample of respondents used in Study 1 comes from the Estonian Biobank cohort, the data for which were collected by the Estonian Genome Centre (EGC) of the University of Tartu (Leitsalu et al., 2014). Participants were recruited on a voluntary basis among the Estonian resident adult population (aged over 18 years). The current number of participants—close to 52 000—represents nearly 5%, of the Estonian adult population. The age structure of the sample is well matched to the age structure of the entire population. Our sample for the current study consisted of 6690

Table 1. Correlations and descriptives of NEO PI-3 item characteristics ($N = 240$ items)

	1.	2.	3.	4.	5.	6.	7.	8.	9.
1. Item variance									
2. Evaluativeness	−.29**								
3. Observability	.12	.16*							
4. Item domain[a]	.12	−.29**	.08						
5. Item position	−.14*	−.18**	.01	.01					
6. Item length	−.10	−.14*	−.39**	.09	.11				
7. Negation[b]	.16*	−.12	−.26**	−.05	−.11	.03			
8. Single-item internal reliability[c]	.08	.19**	.33**	.04	−.08	−.21**	−.20**		
9. Self–other agreement	.63**	−.12	.43**	.35**	−.03	−.22**	−.13*	.36**	
	1.	2.	3.	4.	5.	6.	7.	8.	9.
M	1.15	.37	4.42	1.60	120.50	44.54	.30	.31	.31
SD	.29	.20	.89	.49	69.43	15.80	.46	.14	.08

*$p < .05$; **$p < .01$;

[a]1 = Neuroticism and Agreeableness, 2 = Extraversion, Openness to Experience, and Conscientiousness;

[b]0 = none, 1 = negation;

[c]Single-item internal reliability is based on the average of two estimates—see text for explanation.

Copyright © 2016 European Association of Personality Psychology

DOI: 10.1002/per

participants; 3345 'targets' who provided self-ratings and 3345 'informants' who provided observer-ratings of the targets (see also Allik, Borkenau, Hřebíčková, Kuppens, & Realo, 2015; Mõttus, Allik, Hřebíčková, Kööts-Ausmees, & Realo, 2016; Realo et al., 2015 for sample descriptions). The 3345 targets (59.3% women) had a mean age of 46.4 years (SD = 17.0, ranging from 18 to 91 years). All participants completed the Estonian version of the NEO PI-3 (McCrae et al., 2005). The targets nominated somebody who knew them well and these people were asked to rate the personality traits of the target using the other-report version of the Estonian NEO PI-3. Of the informants, 2331 were women (71.1%) and 948 were men (66 did not report their gender). The mean age of the informants was 41.8 (SD = 15.9) years. The informants had known the participant on average for 23.2 years (SD = 15.1). About 47% of the informants were spouses or partners of the participant, 17% were parents, 16% were friends, 7% were children or grandchildren, 6% were brothers or sisters, and 6% were other relatives or acquaintances.

Instruments

As noted above, the items of the Estonian version of the NEO PI-3 (McCrae et al., 2005), which is a slightly modified version of the NEO PI-R (Costa & McCrae, 1992; Kallasmaa, Allik, Realo, & McCrae, 2000), constitute the sample of items on which we conducted the analyses. Like the original NEO PI-R, the NEO PI-3 has 240 items that measure 30 personality facets, which are grouped into the five FFM domains—Neuroticism (N), Extraversion (E), Openness to Experience (O), Agreeableness (A), and Conscientiousness (C)—such that each domain score is a composite of six facet scores. Except for evaluativeness and observability noted below, items were answered on a 0 (strongly disagree) to 4 (strongly agree) rating scale. The NEO PI-R/NEO PI-3 has excellent psychometric properties in a wide range of countries (De Fruyt, De Bolle, McCrae, Terracciano, & Costa, 2009), including Estonia. The Cronbach alphas of the NEO PI-3 domain scales for self- and other-ratings were .93/.93 (N), .93/.93 (E), .90.89 (O), .87/.92 (A), and .91/.94 (C), respectively. The convergent correlations between self- and other-ratings (i.e. self–other agreement) were .53 (N), .63 (E), .61 (O), .48 (A), and .51 (C).

Item variance. The item variances of self- and other-ratings correlated .86 ($p < .01$) and thus we obtained for each of the 240 NEO PI-3 items a composite item variance variable by averaging the item variances of self- and other-ratings. The mean composite item variance was 1.15 (SD = 0.29).

Evaluativeness. The evaluativeness ratings of the Estonian NEO PI-3 were taken from a study by Mõttus, McCrae, Allik, and Realo (2014). Nine judges rated the items in terms of their social desirability using the following instruction: 'The descriptive characteristics of people often contain an evaluative component. Some characteristics are considered very important for gaining social approval, whereas other characteristics are not approved at all. For each item, please indicate how helpful agreeing with it would be for gaining others' approval.'[5] The ratings were provided on a 7-point scale (1 = not helpful at all to 7 = very helpful). The inter-rater agreement was high (ICC(3, k) = .93). For each item, the mean desirability score was calculated. In agreement with Study 2, we conceptualized scores lower than the midpoint (in this case, '4') as indicating that the characteristic described in the item is met with social disapproval and scores higher than the midpoint as indicating that the characteristic in question is met with social approval. Because scores with higher deviation from the midpoint are less evaluative neutral and because deviations from evaluative neutrality have been found to affect self–other agreement (John & Robins, 1993), we decided to centre and transform the 240 item scores to an absolute (0 to 1) evaluativeness scale using the following transformation: new score = |(old score − 4)/3)|.[6] The mean of the evaluativeness scale was .37, with a standard deviation of .20.

Observability and item domain. The same nine judges also rated the items in terms of their observability using the following instruction: 'Some aspects of personality are easy to judge by external observers, whereas some aspects may be judgeable only by people themselves. For each item, please indicate how easy it would be for an external observer to decide if it describes the person being rated.' The items were rated on a 7-point (very difficult to very easy) scale and the inter-rater agreement (ICC(3, k)) was .84. Similar to evaluativeness, an average was computed for each item. The mean of these 240 observability scores was 4.42 with a standard deviation of .89. Apart from observability scores, to align this study to Study 2, we combined Neuroticism and Agreeableness in one domain (coded '1'), which conforms to the HEXACO Altruism factors, and Extraversion, Conscientiousness, and Openness to Experience in another domain (coded '2'), which conforms to the HEXACO Engagement factors. A t-test showed that the 'Altruism' domain items had significant lower self–other agreement than the 'Engagement' domain items ($t(238) = 5.73$, $p < .01$, $d = 0.76$) and that indeed Agreeableness and Neuroticism items had—on average—the lowest mean self–other agreement correlations of all NEO PI-3 domain scales.[7]

Item position, item length, and negation. Item position was derived from its position in the questionnaire. The 240 items of the NEO PI-3 were presented in a fixed order and thus item position was equal to the item number in the questionnaire. Item length was obtained by counting all characters (excluding spaces and periods but including characters such as commas, brackets, and quotation marks)

[5]As in Study 2, we measured evaluativeness using a general social desirability instruction. Future studies might like to investigate whether instructions with different social desirability 'themes' (e.g. themes that align with engagement/agency or with altruism/communion (cf. Wiggins, 2003) or with each of the five or six personality factors) may yield stronger results for items associated with these themes.

[6]We transformed the scores by dividing by '3' in order to have similar metric as in Study 2. This transformation did not affect the results of the correlational and path analyses.

[7]Note, however, that the *domain-level* self–other agreement of Neuroticism was somewhat higher than Conscientiousness (see text above).

Copyright © 2016 European Association of Personality Psychology

DOI: 10.1002/per

in the self- and the observer-version. The item lengths of the two versions were averaged to arrive at the final item length variable used in this study. The average item length of the Estonian version of the NEO PI-3 was 44.5 characters, ranging from 15 to 108 characters (*SD* = 15.8). Negation was obtained by coding items that contained a negation (e.g. separate words such as 'not,' 'no,' 'nothing,' 'never,' or 'less,' but also pre- or suffixes such as 'dis-,' 'un-,' or '-less') as '1' and items without a negation as '0.' In total, there were 71 items (29.6%) in the Estonian version of the NEO PI-3 that contained a negation.

The criteria variables and sensitivity analysis. The two criteria variables in Study 1 were single-item internal reliability and self–other agreement. To obtain single-item internal reliability, we calculated the mean of two indices. The first index was based on the facet-level single-item internal reliability, for which we explained the calculation in the introduction (e.g. Wanous & Reichers, 1996). The second index was based on the communality of the item with its facet when using one-factor principal axis factoring (Wanous & Hudy, 2001). Across the 240 items, these two indices were very highly correlated (r = .99, $p < .01$) in both self- and other-ratings. The two indices were averaged separately across self- and other-ratings (after r-to-z conversion) and then combined into one single-item internal reliability estimate. The mean single-item internal reliability was .31 (*SD* = .14), with a range of .01 to .67. Self–other agreement was obtained by correlating self- and other-ratings on each of the 240 NEO items. The mean self–other agreement was .31 (*SD* = .08), with a range of .12 to .54. In contrast to Study 2, no retest reliability data was available for the NEO PI-3.

With a sample of 240 items and a statistical power of 80%, sensitivity analyses estimated the smallest effect that correlational analyses will be able to detect with $p < .05$ to be ρ = .16, a small to medium effect size. The k = 240 item-level NEO PI-3 data is made available through the Open Science Framework (De Vries, Realo, & Allik, 2016).

STUDY 1 RESULTS

As shown in Table 1, there were some differences in the way the predictor variables correlated with single-item internal reliability and self–other agreement. The most important correlates of single-item internal reliability (in order of correlation magnitude) were observability, item length, negation, and evaluativeness. The most important correlates of self–other agreement were item variance, observability, item domain, and item length. On the whole, self–other agreement had a stronger correlation with the predictor variables than single-item internal reliability. Two noteworthy findings were (i) that item position did not have a significant relation with single-item internal reliability and self–other agreement, although the (weak) correlations were in the expected direction, with lower single-item internal reliability and self–other agreement when variables were positioned at the end of the questionnaire; and (ii) that observability was unrelated to item domain, although again the weak correlation was in the right direction, with somewhat higher observability for engagement items.

The relations between item variance, evaluativeness, observability, item domain, item position, item length, and negation on the one hand, and the two criteria variables on the other hand were tested using AMOS 21.0 (Arbuckle, 2011). As explained in the introduction, we modelled item variance as a mediator in the relation between evaluativeness and the two criteria variables. Consequently, in the first model we ran, we only freed the path coefficient between evaluativeness and item variance, but not between the other item characteristics and item variance. We did, however, allow all of the item characteristics to be associated with the two criteria variables, that is, single-item internal reliability and self–other agreement. Furthermore, all covariances between the exogenous variables were allowed to be freely estimated, as well as the covariance between the error terms associated with single-item internal reliability and self–other agreement. This first model did not fit optimally, with $\chi^2(5) = 25.97$, $p < .01$; CFI = .95; and RMSEA = .13, p-close < .01. Modification indices showed that three other item characteristics, that is, observability, item position, and negation, were significantly related to item variance. Theoretically, a model which includes these three paths may be defended based on the following observations: (i) *observability* may increase the amount of trait and specific item variance, which in turn may be related to self–other agreement (Watson et al., 2000); (ii) items positioned at the end of a questionnaire may invite more straightlining, and thus *item position* may be associated with a reduction in item variance; and (iii) *negations* may actually result in an increase in item variance because error variance (which may decrease self–other agreement) is added to trait, method, and specific variance. To probe these possible relations, we conducted an exploratory analysis in which we freed these three paths. The resulting model fit the data much better, with $\chi^2(2) = 1.02$, p = 60; CFI = 1.00; RMSEA = .00, p-close = .75, and thus we decided to retain it for further analyses.

We subsequently checked whether a model with or without a direct effect of evaluativeness on self–other agreement had a better fit. Model comparison fit indices ($\Delta\chi^2(1) = 2.87$, p = .09) indicated that a model which excluded the direct path from evaluativeness to self–other agreement did not have a significant worse fit; thus a model without this path was more parsimonious. The model without a direct effect of evaluativeness on self–other agreement fitted well, with $\chi^2(3) = 3.90$, p = .27; CFI = 1.00; RMSEA = .04, p-close = .51, and is shown in Figure 1 and Table 2. We used the bootstrap procedure in AMOS with 5000 samples and a 95% bias-corrected Confidence Interval (CI) to check whether the indirect effect of evaluativeness on self–other agreement through item variance was significant and found evaluativeness had a significant standardized indirect negative relation with self–other agreement (γ = −.20; CI = −.13, −.28).[8]

[8]The indirect effect of evaluativeness on self–other agreement when the direct effect of evaluativeness was also included, was also significant and even slightly stronger, that is, γ = −.21 (CI = −.14, −.29).

Copyright © 2016 European Association of Personality Psychology

DOI: 10.1002/per

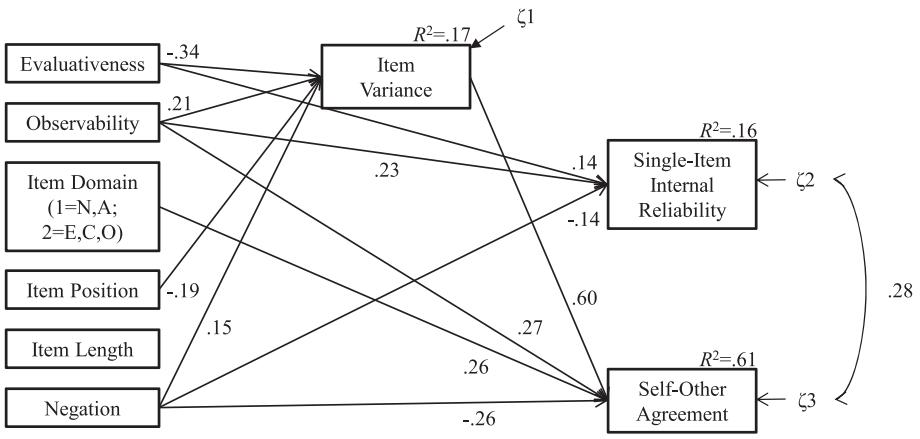

Figure 1. The relations of NEO PI-3 item characteristics to item variance, single-item internal reliability, and self–other agreement; Model fit: $\chi^2(3) = 3.90$, $p = .27$; CFI = 1.00; RMSEA = .04. For clarity, all nonsignificant paths and covariances between the exogenous variables are omitted from the figure. All unstandardized and standardized paths, including the nonsignificant ones, are included in Table 2.

Table 2. Unstandardized and standardized path coefficients and effect sizes of the model in Figure 1 (N = 240 NEO PI-3 items)

	Unstandardized path coefficients (S.E.)		Standardized path coefficients
Direct effects			
Evaluativeness → Item variance	−.480	(.087)	−.34**
Observability → Item variance	.069	(.020)	.21**
Item position → Item variance	−.001	(.000)	−.19**
Negation → Item variance	.095	(.039)	.15*
Evaluativeness → Single-item internal reliability	.092	(.044)	.14*
Observability → Single-item internal reliability	.034	(.011)	.23**
Item domain → Single-item internal reliability	.015	(.018)	.06
Item position → Single-item internal reliability	.000	(.000)	−.05
Item length → Single-item internal reliability	−.001	(.001)	−.09
Negation → Single-item internal reliability	−.04	(.019)	−.14*
Observability → Self–other agreement	.024	(.004)	.27**
Item domain → Self–other agreement	.041	(.007)	.26**
Item position → Self–other agreement	.000	(.000)	.05
Item length → Self–other agreement	.000	(.000)	−.08
Negation → Self–other agreement	−.024	(.007)	−.14**
Item variance → Single-item internal reliability	.043	(.031)	.09
Item variance → Self–other agreement	.161	(.011)	.60**
Indirect effects			
Evaluativeness → Single-item internal reliability	−.021	(.015)	−.03
Evaluativeness → Self–other agreement	−.077	(.015)	−.20**

*$p < .05$; **$p < .01$.

The model in Figure 1 shows that observability, evaluativeness, and negation were related to single-item internal reliability, explaining 16% of its variance, but that except for item position and item length, all other predictor variables, either directly and/or indirectly through item variance, were related to self–other agreement, explaining a substantial 61% of its variance. To check whether the indirect effects of observability and negation on self–other agreement through item variance were significant, we used the same bootstrap procedure in AMOS and found significant standardized indirect relations between observability and self–other agreement, $\gamma = .13$ (CI = .06, .19), between item position and self–other agreement, $\gamma = -.11$ (CI = −.04, −.20), and between negation and self–other agreement $\gamma = .09$ (CI = .02, .17).

STUDY 1 CONCLUSION AND DISCUSSION

Study 1 shows that there is only some overlap in item characteristics that predict single-item internal reliability and self–other agreement. Compared to self–other agreement, single-item internal reliability is relatively weakly predicted by three item characteristics, that is, by evaluativeness, observability, and negation. In contrast, self–other agreement is very strongly predicted by item variance and item domain, both directly and indirectly by observability and negation, and indirectly by evaluativeness. Of seven item characteristics, item variance is by far the most important determinant of self–other agreement and has little impact on single-item internal reliability. Evaluativeness had a negative relation with self–other agreement through item variance and a

Copyright © 2016 European Association of Personality Psychology

DOI: 10.1002/per

positive direct relation with single-item internal reliability. That is, higher evaluativeness seems to be associated with lower item variance, and low variance, in turn, has a negative impact on self–other agreement. Whereas the indirect effect of evaluativeness on self–other agreement was negative, the direct effect of evaluativeness on single-item internal reliability was positive. Consequently, high levels of reliability may mask the unwanted effects of high levels of evaluativeness.

The effects of observability appear to be positive for both single-item internal reliability and self–other agreement. Consequently, with higher levels of observability of an item, higher levels of single-item internal reliability and self–other agreement can be expected. A surprise of this study was, however, that observability was unrelated to item domain. According to scholars, traits such as Extraversion are more visible than traits such as Neuroticism (Funder & Dobroth, 1987; John & Robins, 1993; Watson et al., 2000), and additional analyses confirmed that there was a significant difference in observability of the five NEO PI-3 traits ($F(4, 235) = 17.28$, $p < .01$), with indeed Extraversion as the most observable trait ($M_E = 5.04$, $SD_E = .73$), followed by Conscientiousness ($M_C = 4.62$, $SD_C = .89$), Neuroticism ($M_N = 4.53$, $SD_N = .75$), Agreeableness ($M_A = 4.14$, $SD_A = .87$), and Openness to Experience ($M_O = 3.79$, $SD_O = .88$). As noted in the methods section, self–other agreement of Openness to Experience was relatively high, whereas its observability was relatively low. Item domain (Altruism versus Engagement) may thus be associated with something else than just observability. We will get back to this point in the general discussion.

The structural equation model did not sustain a significant relation between item length and the criteria, possibly because its effects were somewhat confounded with those of observability, with which it shared a negative relation ($r = -.39$, $p < .01$). Although item position did have a negative indirect effect on self–other agreement through item variance, because this effect was compensated with a positive (nonsignificant) effect, its total effect was not significantly different from zero, $\gamma = -.07$ (CI = −.17, +.04). Similarly, the positive indirect effect of negation on self–other agreement was compensated with a significant negative direct effect, resulting in a nonsignificant total effect of negation on self–other agreement, $\gamma = -.05$ (CI = −.17, +.07). Consequently, the total effects of item position, item length, and negation on single-item internal reliability and self–other agreement were weak or absent.

Furthermore, it should be noted that some suppressor effects seemed to occur in the structural equation model. That is, the zero-order relations of evaluativeness, observability, and item position with item variance were lower in the correlation matrix than in the structural equation model and whereas the correlation between evaluativeness and self–other agreement was nonsignificant in the correlation matrix, the total standardized effect was significant in the structural equation model, even when the direct effect of evaluativeness on self–other agreement was included. Consequently, although the results seem to offer by-and-large support for our expectations with respect to the relevance (item variance and evaluativeness) and availability (observability and item domain) variables but not for our detection (item position) and utilization (item length and negation) variables, more research is clearly needed.

STUDY 2

Study 2 is a replication and extension of Study 1 using a different sample, a different questionnaire, and an additional criterion variable (i.e. retest reliability). First of all, instead of the NEO PI-3, the HEXACO-PI-R (Ashton & Lee, 2008; De Vries et al., 2009; Lee & Ashton, 2004) was used in Study 2. Second, we used an additional sample from the Netherlands that provided test–retest data to compare the effects of the predictors on reliability and self–other agreement with their effects on retest reliability. This addition may be deemed important, because recent studies have suggested that retest reliability may actually be a more useful indicator of reliability than internal reliability estimates, such as Cronbach's alpha (De Vries, 2013; McCrae et al., 2011). That is, we predicted that the pattern of correlations of item variance, evaluativeness, observability, item domain, item length, negation, and item position will be highly similar for retest reliability and self–other agreement but different for single-item internal reliability.

Method

Sample and procedure

The effective sample for the analyses of Study 2 consisted of the entire set of 200 items of the HEXACO-PI-R, described under 'Instruments.' The main data on these 200 items were derived from a sample of respondents, consisting of 4396 participants, 2198 of whom were Dutch first year personality psychology students (81.7% women; $M_{age} = 20.2$; $SD_{age} = 2.8$). The psychology students filled out a number of personality inventories as part of their coursework and approached a well-acquainted other to obtain informant reports on their personality. The sample of 2198 well-acquainted observers (63.2% women; $M_{age} = 26.8$; $SD_{age} = 12.3$) consisted of 41.2% friends, 35.9% family members, and 22.9% intimate partners, who on average knew the focal person for 10.6 years ($SD = 8.0$). Both self- and other-ratings were used by another student to write a personality report about the focal person. The dataset, or parts of it, has been used in other studies (e.g. Allik, De Vries, & Realo, 2016; De Vries et al., 2009; De Vries, Wawoe, & Holtrop, 2016, Studies 1 and 4), but the dataset has not been used for the present purpose.

Instruments

HEXACO-PI-R

The six domain scales that form the HEXACO acronym are represented each by 32 items and eight items measure the interstitial Altruism facet. The total HEXACO Personality Inventory-Revised thus consists of 200 items. The Proactivity interstitial facet, that has recently been added to the Dutch version (De Vries et al., 2016), was not included

Copyright © 2016 European Association of Personality Psychology

DOI: 10.1002/per

in this study. All items were answered on a 1 (strongly disagree) to 5 (strongly agree) rating scale. The domain-level alpha reliabilities of the self/other versions in this sample were respectively .90/.91 (H), .90/.89 (E), .91,.92 (X), .89/.91 (A), .90/.91 (C), and .89/.89 (O) and the convergent correlations between self- and other-ratings (i.e. self–other agreement) were .49 (H), .63 (E), .64 (X), .53 (A), .63 (C), and .64 (O).

Item variance. Similar to Study 1, the correlation between self- and other-rated item variance was high ($r = .88$, $p < .01$) and thus self- and other-rated item variances were averaged to obtain the composite item variance variable. The mean composite item variance was .99 ($SD = 0.23$).

Evaluativeness. Six psychologists (50% women; $M_{age} = 33.3$, $SD_{age} = 10.4$) rated the social desirability (i.e. 'the extent to which the individual characteristic described in the item is viewed as socially desirable by people in general') of the 200 HEXACO items on a five-point rating scale (1 = socially very undesirable to 5 = socially very desirable). The inter-rater agreement was high (ICC(3, k) = .89), and the items were subsequently centred and transformed to an absolute (0 to 1) Evaluativeness scale similar to Study 1 using the following transformation: new score = |(old score − 3)/2)|. The scale mean was .33, with a standard deviation of .21.

Observability and item domain. A different set of eight psychologists (37.5% women; $M_{age} = 34.1$, $SD_{age} = 7.5$) rated the observability (i.e. 'the extent to which the individual characteristic described in the item is difficult or easy to observe for an external observer') of the 200 HEXACO items on a five-point (1 = very difficult to observe to 5 = very easy to observe) scale. The inter-rater agreement was adequate (ICC(3, k) = .80), and the scale mean was 2.90, with a standard deviation of 0.53. Furthermore, for item domain, we coded items belonging to the Altruism domains (Honesty–Humility, Emotionality, and Agreeableness) as '1,' and items belonging to the Engagement domains (Extraversion, Conscientiousness, and Openness to Experience) as '2.'

Item position, item length, and negation. In line with Study 1, we included the following additional predictor variables: item position, item length, and negation. The item position variable was based on the position of the variable in the questionnaire. The items of the HEXACO questionnaire were presented in a fixed order and thus item position was equal to the item number in the questionnaire. Per item, all characters—including quotation marks, commas, brackets, etc. but excluding periods and spaces—were used for the item length variable, which was obtained by averaging the item lengths across self- and observer versions. The mean item length was 60.6 characters ($SD = 16.9$) with a range of 14.5 to 109.5 characters. The negation variable was based on the presence of negations (e.g. not, no, nothing, never, less, dis-, un-, -less, etc…) in the item. Items that contained a negation were coded as '1,' items without a negation were coded as '0.' As noted in the introduction, there were 67 HEXACO-PI-R items (33.5%) with negations.

The criteria variables and sensitivity analysis. The three criteria variables were single-item internal reliability, retest reliability, and self–other agreement. Single-item internal reliability was operationalized using the same method as in Study 1 (i.e. by averaging facet-level single-item internal reliabilities and communalities based on a principal axis factoring analysis). Again, correlations between the two indices were very high, reaching almost parity for both self- ($r = .996$, $p < .01$) and other-ratings ($r = .997$, $p < .01$).

Retest reliability was based on an earlier sample of $N = 188$ students (85.1% women; $M_{age} = 19.7$, $SD_{age} = 2.3$ at T1), who completed the unrevised version of the HEXACO Personality Inventory (De Vries, Lee, & Ashton, 2008; Lee & Ashton, 2004). The students completed the HEXACO-PI two times in 2006; the first time during a first year personality psychology course (T1), and a second time after seven months during a second year methodology course (T2). The retest correlations between the six domain scales ranged from .79 (Agreeableness) to .90 (Openness to Experience), with a mean of .85. To check for potential score manipulation by students who had obtained feedback on their HEXACO personality profile before T2, we conducted six pairwise *t*-tests. All in all, the numbers were not suggestive of large scale score manipulations by psychology students. That is, none of the means on the three Altruism dimensions were significantly different, that is, Honesty–Humility ($M_{t1} = 3.60$, $SD_{t1} = .51$; $M_{t2} = 3.60$, $SD_{t2} = .53$; $t(df = 187) = .36$, $p = .72$, $d = 0.03$), Emotionality ($M_{t1} = 3.41$, $SD_{t1} = .47$; $M_{t2} = 3.42$, $SD_{t2} = .45$; $t(df = 187) = -.80$, $p = .43$, $d = -0.06$), and Agreeableness ($M_{t1} = 3.01$, $SD_{t1} = .48$; $M_{t2} = 2.97$, $SD_{t2} = .45$; $t(df = 187) = 1.90$, $p = .06$, $d = 0.14$) and although the means on three Engagement dimensions were significantly different, that is, Extraversion ($M_{t1} = 3.31$, $SD_{t1} = .50$; $M_{t2} = 3.37$, $SD_{t2} = .49$; $t(df = 187) = -2.93$, $p < .01$, $d = -0.21$), Conscientiousness ($M_{t1} = 3.30$, $SD_{t1} = .46$; $M_{t2} = 3.38$, $SD_{t2} = .45$; $t(df = 187) = -4.47$, $p < .01$, $d = -0.33$), and Openness to Experience ($M_{t1} = 3.29$, $SD_{t1} = .54$; $M_{t2} = 3.35$, $SD_{t2} = .51$; $t(df = 187) = -3.24$, $p < .01$, $d = -0.24$), in absolute terms, the mean differences were not very large.

The only difference between the HEXACO-PI and the HEXACO-PI-R is that the former contained an Expressiveness facet which was replaced by a Social Self-Esteem facet in the HEXACO-PI-R. Consequently, eight items in the retest sample were different from those used in the remainder of our study. The values of the eight Social Self-Esteem items were imputed with the EM algorithm using the self- and other item-rest correlations as predictors. When using the imputed values, all correlations in Table 3 remained virtually unchanged.

Self–other agreement was obtained by correlating self- and other-ratings on each of the 200 HEXACO items. The average self–other agreement of the items was .30 (ranging from .09 to .56), with a standard deviation of .09.

With a sample of 200 items and a statistical power of 80%, sensitivity analyses estimated the smallest effect that correlational analyses will be able to detect with $p < .05$ to be $\rho = .17$, a small to medium effect size. The $k = 200$ item-level HEXACO-PI-R data is made available through the Open Science Framework (De Vries et al., 2016).

Copyright © 2016 European Association of Personality Psychology

DOI: 10.1002/per

Table 3. Correlations and descriptives of HEXACO-PI-R item characteristics ($N = 200$ items)

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
1. Item variance										
2. Evaluativeness	−.24**									
3. Observability	−.08	−.09								
4. Item domain[a]	.13	−.17*	.02							
5. Item position	−.07	.02	−.14*	−.05						
6. Item length	.06	.03	−.16*	−.04	.02					
7. Negation[b]	−.01	.07	−.12	−.11	−.05	−.01				
8. Single-item internal reliability[c]	.26**	.04	.06	.07	−.04	.01	.02			
9. Retest reliability	.51**	−.06	.04	.34**	−.12	−.30**	−.16*	.31**		
10. Self–other agreement	.63**	−.21**	.16*	.43**	−.08	−.23**	−.14	.43**	.69**	
	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
M	.99	.33	2.90	1.48	100.50	60.60	.33	.35	.56	.30
SD	.23	.21	.53	.50	57.88	16.90	.47	.12	.10	.09

*$p < .05$; **$p < .01$;
[a]1 = Altruism, 2 = Engagement;
[b]0 = none, 1 = negation;
[c]Single-item internal reliability is based on the average of two estimates—see text for explanation.

STUDY 2 RESULTS

There were again strong differences in correlates of single-item internal reliability on the one hand and retest reliability and self–other agreement on the other. The only significant correlate of single-item internal reliability in the HEXACO sample was item variance, whereas retest reliability and self–other agreement were strongly and consistently related to item variance, item domain, and (negatively) to item length. In addition, self–other agreement was also negatively related to evaluativeness and positively to observability. Retest reliability was also more strongly related to self–other agreement ($r = .69$, $p < .01$) than it was to single-item internal reliability ($r = .31$, $p < .01$) or than single-item internal reliability was to self–other agreement ($r = .43$, $p < .01$). Item position was unrelated to the three criteria variables, although all three correlations were in the expected (negative) direction. Again, as in Study 1, observability was unrelated to item domain.

As in Study 1, our first structural equations model in AMOS 21.0 (Arbuckle, 2011) included item variance as a mediator of the relation between evaluativeness and the three criteria variables but not as a mediator between the other item characteristics and the three criteria variables. Similar to Study 1, all exogenous variables were allowed to covary as were the error terms associated with the three criteria variables. This first model fit the data really well, with $\chi^2(5) = 6.17$, $p = .29$; CFI = 1.00; RMSEA = .03, p-close = .55. To compare our model with the exploratory model in Study 1, we subsequently freed up the paths from observability, item position, and negation to item variance. The model which included these paths did not significantly improve model fit (e.g. $\Delta\chi^2(3) = 3.82$, $p = .28$) and all of the three paths were not significantly different from zero. Consequently, we decided to retain the original model.

Subsequently, we checked whether a model with or without a direct effect of evaluativeness on self–other agreement had a better fit. Model comparison fit indices ($\Delta\chi^2(1) = 0.09$, $p = .77$) indicated that a model which excluded the direct path was more parsimonious. This final model, which did not have a direct effect of evaluativeness on self–other agreement, also fitted well, with $\chi^2(6) = 6.25$, $p = .40$; CFI = 1.00; RMSEA = .02, p-close = .67, and is shown in Figure 2 and Table 4.

The item characteristics explained 9% of the variance in single-item internal reliability, 47% of the variance in retest reliability, and 62% of the variance in self–other agreement. Using a bootstrap procedure in AMOS with 5000 samples and a 95% bias-corrected Confidence Interval (CI), the standardized parameters of the indirect effects indicated that evaluativeness had a significant standardized indirect relation with single-item internal reliability ($\gamma = -.07$; CI = −.03, −.13), retest reliability ($\gamma = -.13$; CI = −.06, −.21), and self–other agreement ($\gamma = -.15$; CI = −.07, −.23).[9] However, evaluativeness only had a significant standardized total effect on self–other agreement ($\gamma = -.15$; CI = −.07, −.23). The standardized total effects of evaluativeness on single-item internal reliability ($\gamma = .05$; CI = −.10, .19) and retest reliability ($\gamma = -.01$; CI = −.13, .12) were nonsignificant.

STUDY 2 CONCLUSION AND DISCUSSION

The main outcome of Study 2 is that, again, item variance is the main predictor of self–other agreement and is more strongly related to self–other agreement than it is related to single-item internal reliability. Interestingly enough, the relations of item variance, item length, and item domain with retest reliability were almost similar to their relations with self–other agreement, indicating that similar processes may be at play in retest and self–other agreement data. That is, just like other-ratings, a person may agree less with him-/herself a second time around when there is less variance in the item, when the item is somewhat longer, and when the questions are about altruism factors instead of engagement factors.

[9]The indirect effect of evaluativeness on self–other agreement was exactly the same when the direct effect was included.

Copyright © 2016 European Association of Personality Psychology

DOI: 10.1002/per

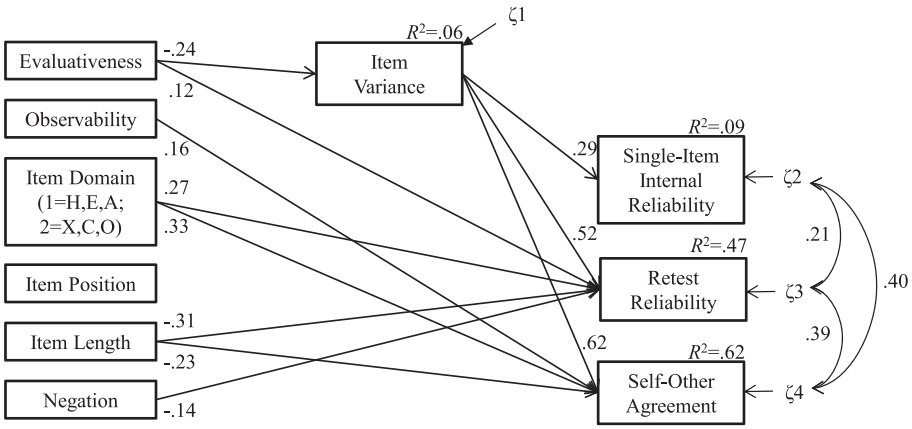

Figure 2. The relations of HEXACO-PI-R item characteristics to item variance, single-item internal reliability, retest reliability, and self–other agreement; Model fit: $\chi^2(6) = 6.25$, $p = .40$; CFI = 1.00; RMSEA = .02. For clarity, all nonsignificant paths and covariances between the exogenous variables are omitted from the figure. All unstandardized and standardized paths, including the nonsignificant ones, are included in Table 4.

Table 4. Unstandardized and standardized path coefficients and effect sizes of the model in Figure 2 (N = 200 HEXACO-PI-R items)

	Unstandardized path coefficients	(S.E.)	Standardized path coefficients
Direct effects			
Evaluativeness → Item variance	−.266	(.075)	−.24**
Evaluativeness → Single-item internal reliability	.065	(.035)	.12
Observability → Single-item internal reliability	.020	(.015)	.09
Item domain → Single-item internal reliability	.012	(.016)	.05
Item position → Single-item internal reliability	.000	(.000)	−.01
Item length → Single-item internal reliability	.000	(.000)	−.02
Negation → Single-item internal reliability	.008	(.017)	.03
Evaluativeness → Retest reliability	.056	(.024)	.12*
Observability → Retest reliability	.002	(.010)	.01
Item domain → Retest reliability	.053	(.011)	.27**
Item position → Retest reliability	.000	(.000)	−.07
Item length → Retest reliability	−.002	(.000)	−.31**
Negation → Retest reliability	−.030	(.011)	−.14**
Observability → Self–other agreement	.029	(.008)	.16**
Item domain → Self–other agreement	.062	(.008)	.33**
Item position → Self–other agreement	.000	(.000)	.01
Item length → Self–other agreement	−.001	(.000)	−.23**
Negation → Self–other agreement	−.016	(.009)	−.08
Item variance → Single-item internal reliability	.145	(.035)	.29**
Item variance → Retest reliability	.226	(.023)	.52**
Item variance → Self–other agreement	.248	(.018)	.62**
Indirect effects			
Evaluativeness → Single-item internal reliability	−.038	(.013)	−.07**
Evaluativeness → Retest reliability	−.060	(.018)	−.13**
Evaluativeness → Self–other agreement	−.066	(.019)	−.15**

*$p < .05$; **$p < .01$.

In line with Study 1, evaluativeness was indirectly—through item variance—negatively related to self–other agreement. Also in line with Study 1, observability and item domain were positively related to self–other agreement but were unrelated to each other. Additional analyses showed that there was a significant difference in observability between the six HEXACO traits ($F(5, 186) = 13.99$, $p < .01$), with—as in Study 1—Extraversion the most observable trait ($M_X = 3.23$, $SD_X = 0.72$), followed by Agreeableness ($M_A = 3.12$, $SD_A = 0.36$), Emotionality ($M_E = 3.08$, $SD_E = 0.39$), Conscientiousness ($M_C = 2.96$, $SD_C = 0.33$), Openness to Experience ($M_O = 2.54$, $SD_O = 0.44$), and Honesty–Humility ($M_H = 2.54$, $SD_H = 0.41$). Again, as in Study 1, self–other agreement of Openness to Experience was relatively high, whereas its observability was relatively low.

As in Study 1, negation and item position had relatively weak or absent relations with the three criteria variables. Item length was somewhat more strongly related to the criteria and —in contrast to Study 1—remained a significant predictor in the SEM. Longer items were associated with somewhat lower levels of retest reliability and self–other agreement than shorter items.

Copyright © 2016 European Association of Personality Psychology

DOI: 10.1002/per

GENERAL DISCUSSION AND CONCLUSIONS

Recent articles by McCrae and colleagues (McCrae, 2015; McCrae et al., 2011) have challenged the widespread assumption that a personality inventory is as good as its internal reliability. Instead, the most important psychometric indices of a scale seem to be high levels of retest reliability and self–other agreement. In previous research (Allik et al., 2010), variance was found to be the most important predictor of self–other agreement at the facet level. In two studies, we replicated this finding at the item level. That is, data obtained from two large samples speaking two very different languages, Dutch (Indo-European) and Estonian (Finno-Ugric), using two principally different questionnaires, the NEO PI-3 and the HEXACO-PI-R, showed that item variance is the most important predictor of self–other agreement and retest reliability and that it is, at the same time, not highly predictive of single-item internal reliability. Furthermore, data from both studies also suggest that item variance mediates the negative relation of evaluativeness with retest reliability and self–other agreement. That is, evaluativeness is an impediment of item variance, and lower levels of item variance may reduce retest reliability and self–other agreement. Apart from item variance and evaluativeness, observability and item domain also seem to play a role, with item content associated with the altruism domain and lower observability being associated with lower self–other agreement.

The results have important implications for the construction of both long and short versions of personality questionnaires, although the implications for short versions are somewhat more pronounced. Based on our findings, when constructing personality questionnaires we would advise scholars to investigate and report for each item its (i) level of self–other agreement; (ii) level of retest reliability; (iii) item variance; (iv) evaluativeness; and (v) observability. Instead of using internal reliability as a screening mechanism, scholars could use expert raters to assess the face validity of an item with its purported construct and assess whether items are not tautological to prevent 'bloated specifics' (Cattell, 1973). For example, in the HEXACO-PI-R, 80 items had a retest reliability $\geq .60$ and there were ample items from each domain with sufficient levels of self–other agreement ($> .30$), item variance (>1.00), low levels of evaluativeness ($< .30$), and high levels of observability (e.g. > 3.00 on a 5-point scale). The exact combination of the above to screen items and optimize predictive validity remains to be investigated, but based on previous research (McCrae et al., 2011), internal reliability (or in our case: single-item internal reliability) does not seem to have much of an impact on predictive validity.

The above suggestions are especially important for short scales, consisting of two to five items. Items for short scales are often selected based on their item-total correlations and internal reliability estimates (e.g. Donnellan, Oswald, Baird, & Lucas, 2006; Rammstedt & John, 2007; Thalmayer et al., 2011) and although it may still be important to obtain short scales that are aligned with the main vector position of longer versions, researchers should probably pay more heed to retest reliability and self–other agreement when selecting items for short scales to optimize their stability and predictive validity. As a case in point, De Vries (2013) showed that a weighted combination of internal reliability, retest reliability, and self–other agreement indices to disattenuate predictive validity of the Brief HEXACO Inventory provided a more accurate estimate of predictive validity than a correction based on separate internal reliability, retest reliability, or self–other agreement indices. Thus, short scales that mainly optimize internal reliability may optimize method variance at the expense of trait and specific variance, sacrificing construct breadth and validity for content homogeneity.

To optimize trait variance relative to other sources of variance, Generalizability Theory (Shavelson, Webb, & Rowley, 1989; Ziegler, Poropat, & Mell, 2014) may offer an especially informative framework to evaluate personality scales. Generalizability Theory extends classical test theory by distinguishing and estimating multiple sources of variance. Optimally, a design inspired by Generalizability Theory would include at least self- and (multiple) other-ratings of a set of items that belong to the same scale on (at least) two different time points, allowing the decomposition of 15 sources of variance, that is, four main effects (that is, persons variance, items variance,[10] raters variance, and time variance), six two-way interactions (for instance, persons × items variance), four three-way interactions (for instance, persons × items × raters variance), and one residual component (that is, persons × items × raters × time + error variance). Especially when comparing longer versions of a questionnaire to shorter versions, Generalizability Theory may inform researchers which combination of items yields the most optimal short scale in terms of absolute and relative person (trait) variance. Additionally, it might show which specific variance components should be reduced. For instance, Ziegler et al. (2014) showed that there was a substantial amount of persons × items variance (i.e. average relative variance of 25%) in both long and short versions of questionnaires based on the NEO PI-R (Costa & McCrae, 1992) and the Big Five Inventory (John, Donahue, & Kentle, 1991), suggesting that different persons might understand items differently.[11] However, although Generalizability Theory may inform researchers on the optimal combination of items in a scale, it does not tell us what item characteristics may yield a better scale.

In line with findings at the facet level (Allik et al., 2010), results of our studies suggest that scholars should especially focus on (between-persons) item variance to optimize retest reliability and self–other agreement. Item variance was substantially more strongly related to self–other agreement and retest reliability than to single-item internal reliability, that is, $r = .63$ with self–other agreement in Study 1 and 2

[10]Note that 'items variance' in Generalizability Theory is different from 'item variance' as conceptualized in our study. The former is the 'between-items' variance, the latter is the 'between-persons' variance in each single item.

[11]Actually, McCrae (2015) equates 'persons × items' variance with item-specific variance (s) and its reduction might have a detrimental effect on predictive validity when broad traits are conceptualized as consisting of trait variance plus item- and facet-specific variance. According to McCrae (2015), item- and facet-specific sources of variance increment the prediction of criteria (cf. a 'union' ($\cup$) perspective).

Copyright © 2016 European Association of Personality Psychology

DOI: 10.1002/per

versus $r = .08$ and $r = .26$ with single-item internal reliability in Study 1 and 2 and $r = .51$ with retest reliability versus $r = .26$ with single-item internal reliability in Study 2, which—using a test of difference in correlated correlations (Meng, Rosenthal, & Rubin, 1992)—were all highly significant (respectively $z = 8.16$ (Cohen's (1988) effect size $q = 0.66$), $z = 5.29$ ($q = 0.48$), and $z = 3.38$ ($q = 0.30$); all p's $< .01$). That is, item variance seems to be a 'sine qua non' of self–other agreement and retest reliability in personality items and seems to be most important for self–other agreement.

But how does one go about writing items with sufficient high levels of item variance? Although observability was related to item variance in Study 1 (but only in the structural equation model), this effect was not replicated in Study 2. A somewhat different wording of the observability instruction in the two studies, or a slightly higher observability of the NEO PI-3 items (even after correction for the use of a different (1–7 versus 1–5) response scale), may be (part of) the cause. Note, however, that the effect of observability on item variance in Study 1 was not very strong in the first place. The only consistent finding from both studies is that evaluativeness is negatively related to item variance, so writing neutral items seems to be an important first recommendation to arrive at high levels of item variance. However, some may object to this recommendation by reasoning that neutral items with high levels of variance are less likely to discriminate between respondents at high levels of a trait. Indeed, research by Suzuki, Samuel, Pahlen, and Krueger (2015) suggests that the inclusion of items with more extreme content may have some advantages in the discrimination of persons at high or low levels of the trait in question. However, other studies (Ashton et al., in press) have shown that scales which contain extreme (socially undesirable) content are much more likely to be saturated with variance because of response biases. Furthermore, nonclinical personality questionnaires have predicted clinical diagnosis nearly as well as clinical scales that were specifically designed to predict these diagnoses (Quirk, Christiansen, Wagner, & McNulty, 2003). Consequently, the added value of discriminating extreme respondents may not weigh up to the negative effects of extreme items in terms of increased response biases and lower levels of self–other agreement.

Item position, item length, and negation did not have a consistent impact on item variance across the two studies. In the NEO PI-R study (Study 1), item position and negation did have a significant (respectively negative and positive) effect on item variance, but this effect was not replicated in the HEXACO-PI-R study (Study 2). One of the reasons for the impact of item position on item variance in Study 1 but not in Study 2 is that the latter study included psychology students who might be more interested in personality and thus more motivated to fill out a long questionnaire. Among other-ratings, who might have been less motivated to fill out the questionnaires, the effect of item position on item variance was stronger than among self-ratings in both studies, that is, $r_s = -.11$ versus $r_o = -.18$ in Study 1 (using test of difference in correlated coefficients $z = 2.12$, $p = .03$, Cohen's $q = 0.07$) and $r_s = .01$ versus $r_o = -.15$ ($z = 4.54$, $p < .01$, $q = 0.16$) in Study 2. Additionally, older respondents may suffer more from loss of concentration at the end of a questionnaire than younger respondents. Indeed, among other-ratings in Study 1, the effect of item position on item variance was most pronounced among older respondents (60 years and older; $n = 529$) when compared to younger respondents (40 years and younger; $n = 1638$), that is, $r_{young} = -.15$ versus $r_{old} = -.21$ (using test of difference in correlated coefficients $z = 2.25$, $p = .02$, Cohen's $q = 0.07$). Consequently, motivation and decline in age-related concentration may indeed play a role and thus randomization of items might help to counter the negative effects of item position on item variance.

Interestingly, the effect of negation on item variance was positive instead of negative in Study 1, although the extra variance might be mainly error variance, as suggested by the significant negative correlation between negation and single-item internal reliability in both studies. Apart from evaluativeness and possibly item position and negation, future research may like to investigate what other factors play a role in item variance.[12] One conclusion might be that there are instances in which people vary consistently and widely in the breadth of behaviours, thoughts, and feelings, and that making people think about these contexts using questionnaire items may provide the highest levels of item variance. Two examples of items from the HEXACO-PI-R with high item variance and high self–other agreement are 'I could let my room get very messy before I would clean it' (mean item variance = 1.67, self–other agreement = .50) and 'Attending a play is not something that I would enjoy' (mean item variance = 1.54, self–other agreement = .53). Both items seem to outline contexts that invite large individual differences in reactions that are relatively easily available, detected, and 'correctly' utilized by targets and their acquaintances. Apart from writing items as evaluatively neutral as possible, this may be the 'unsolved mystery' (McCrae, 2015) of writing better items—that is, outlining the contexts in which trait expressions vary consistently and widely.

Apart from item variance and evaluativeness, observability and item domain were the two other most important predictors of self–other agreement and, to a lesser extent, retest reliability. An unexpected finding from both studies was that observability and item domain were unrelated. Remember that item domain made a distinction between items that were associated with the Engagement domain (i.e. Extraversion, Conscientiousness, and Openness to Experience for both the NEO PI-3 and the HEXACO-PI-R) and items that were associated with the Altruism domain (i.e. Neuroticism and Agreeableness in the NEO PI-3 and Emotionality, Agreeableness, and

[12]A plausible other factor, suggested by an anonymous reviewer, is item complexity, and, apart from item length, maximum word length may be another proxy for item complexity. We tested this assumption by counting the characters of the longest word in each sentence (i.e. 'maximum word length'), but neither in the NEO PI-3 nor in the HEXACO-PI-R data, maximum word length was related to single-item internal reliability, retest reliability, and/or self–other agreement.

Copyright © 2016 European Association of Personality Psychology

DOI: 10.1002/per

Honesty–Humility in the HEXACO-PI-R). We expected items from the Engagement domain to be more visible than items from the Altruism domain, but this expectation was not confirmed. This lack of relation seemed to be mainly because of Openness to Experience, which was characterized by relatively low levels of observability in both studies. It may thus be that the expression of Openness to Experience (i.e. somebody's creativity, imagination, and aesthetic interest and his/her openness to new ideas, unconventional people, values, and feelings) is less directly observable, whereas trait levels of Openness to Experience are readily deduced from interactions, the content and style of communication, and from behavioural traces (e.g. 'rooms with a cue'; Gosling, Ko, Mannarelli, & Morris, 2002).

However, while unrelated to observability, item domain was still—and substantially—related to self–other agreement and retest reliability. In general, Altruism domain items had lower levels of self–other agreement and retest reliability than Engagement domain items. If not related to observability, what is the reason for this finding? In both studies we found that item domain was related to evaluativeness, but this did not fully explain the relations with self–other agreement and retest reliability. One possible explanation for the effects of item domain may be the extent to which behaviours in the domain tend to fluctuate across time, place, or different interaction partners (cf. 'occasion specificity'; Deinzer et al., 1995). Experience sampling does not provide much evidence, however, that within-person variability is higher for Altruism domain variables than for Engagement domain variables; in fact, research on the Big Five suggests that within-person variability is actually somewhat greater for Engagement domain variables than for Altruism domain variables (Fleeson, 2001; see Figure 2). Still, it may be true that different interaction partners cause people to act differently more on Altruism domains than on Engagement domains. There is some evidence for this stance when inspecting self–other agreement using raters from different contexts. In the NEO PI, mean self–other agreement across three different contexts (parents, college, and hometown) was .46 (Extraversion), .45 (Openness to Experience), and .38 (Conscientiousness) for the Engagement domains and .38 (Agreeableness) and .36 (Neuroticism) for the Altruism domains (Funder et al., 1995; numbers obtained from their Table 1 using *r*-to-*z* transformed correlations). That is, average self–other agreement among different raters appears somewhat higher for the Engagement domains than for the Altruism domains. Thus, specific systematic variations in altruism-related behaviours associated with interaction partners may account for the relation observed between item domain and both self–other agreement and retest reliability. Of course, more research is needed to investigate this proposition.

The findings with respect to item length, negation, and item position were less strong. In the two correlation matrices, item length was found to be negatively related to self–other agreement, retest reliability, and (only in Study 1) single-item internal reliability. Although the effects of item length only held in the SEM in Study 2,[13] the results do seem to suggest that shorter items may generally be favoured over longer items, as has been suggested by Hendriks (1997); Hendriks et al. (1999), and Mõttus et al. (2006). Furthermore, in line with Marsh (1986); Hofstee (1991), and McCrae et al. (2011), who suggested that negations in items lead to ambiguity or lack of comprehensibility, some relations between negations and single-item internal reliability, retest reliability, and self–other agreement were found, although these effects—when present—were relatively weak. Based on these findings, researchers are advised to be prudent when using negations in questionnaire items. Finally, with respect to item position, although—as noted above—item effect was related to item variance among older other-raters, no effect of item position was found on single-item internal reliability, retest reliability, and self–other agreement. Although the items were all in fixed order, and thus item position may be confounded with the content of the item, it is highly unlikely that this will have affected the effects of item position because both questionnaires alternate domains and facets in the presentation of the items.

Although a substantial amount of variance was explained by the seven variables included in our research, there might be other item characteristics that play a role in the explanation of single-item internal reliability, retest reliability, and self–other agreement. For instance, prototypicality (i.e. representativeness) of the item for the domain it belongs to may be an important factor in the explanation of single-item internal reliability and might have less an effect on retest reliability and self–other agreement. Expert ratings of complexity of an item may also yield stronger effects than the more indirect measures of item length and negation. Future research may further investigate these and possible other item characteristics to explain single-item internal reliability, retest reliability, and self–other agreement. Furthermore, future studies may also like to expand on the number of occasions on which respondents are asked to fill out the personality questionnaires. In our research, retest reliability was only assessed in Study 2 using the HEXACO-PI-R and then only after seven months. Researchers have recommended to use a shorter timeframe to assess retest reliability (e.g. maximally a few weeks), and to separate short-term stability (which has been called the *dependability* of a measure) from long-term stability (which has been called the *systematic* stability of a measure, Wood & Wortman, 2012). Items with extreme means (and thus lower variance) have been shown to be infused with (unstable) transient errors, which impacted dependability, but not systematic stability (Wood & Wortman, 2012). Thus, measuring personality items at multiple time-points seems to be important in separating

[13] Although the Estonian NEO PI-3 and the Dutch HEXACO-PI-R had comparable self–other agreement (e.g. .31 (*SD* = .08) versus .30 (*SD* = .09)), the average item length of the Estonian NEO PI-3 was close to a standard deviation lower than the average item length of the Dutch HEXACO-PI-R (e.g. 44.5 (*SD* = 15.8) versus 60.6 (*SD* = 16.9)), which may suggest that further reductions of item length may not have much effect on self–other agreement or may even—as has been suggested in the introduction—reduce self–other agreement.

Copyright © 2016 European Association of Personality Psychology

DOI: 10.1002/per

stable from unstable parts (e.g. 'true' trait changes) of the different systematic variance components.

To conclude, this research shows that single-item internal reliability on the one hand and self–other agreement and retest reliability on the other are differentially predicted by item variance, evaluativeness, observability, and item domain. Awareness of the importance of these item characteristics, and the importance of obtaining sufficient levels of retest reliability and self–other agreement in order to secure predictive validity, may constitute the necessary ingredients to obtain the best possible personality measurement tools.

ACKNOWLEDGEMENTS

Preparation of this manuscript was supported by the University of Tartu (SP1GVARENG) and by institutional research funding (IUT2-13) from the Estonian Ministry of Education and Science. We are grateful to the Estonian Genome Centre of the University of Tartu and its director, Andres Metspalu, for their help in collecting the data and for their kind permission to use the data in the current study. Anu Realo was a Visiting Professor in the Health, Medical, and Neuropsychology Unit of the Faculty of Social and Behavioral Sciences at Leiden University during the writing of this article.

REFERENCES

Allik, J., Borkenau, P., Hřebíčková, M., Kuppens, P., & Realo, A. (2015). How are personality trait and profile agreement related? *Frontiers in Psychology*, *6*, 785. doi:10.3389/fpsyg.2015.00785.

Allik, J., De Vries, R. E., & Realo, A. (2016). Why are moderators of self–other agreement difficult to establish? *Journal of Research in Personality*, *63*, 72–83. doi:10.1016/j.jrp.2016.05.013.

Allik, J., Realo, A., Mõttus, R., Esko, T., Pullat, J., & Metspalu, A. (2010). Variance determines self-observer agreement on the Big Five personality traits. *Journal of Research in Personality*, *44*(4), 421–426.

Arbuckle, J. L. (2011). *IBM SPSS Amos 20 user's guide*. Armonk, NY: *IBM*.

Ashton, M. C., De Vries, R. E., & Lee, K. (in press). Trait variance and response style variance in the scales of the Personality Inventory for DSM-5 (PID-5). *Journal of Personality Assessment*. doi:10.1080/00223891.2016.1208210.

Ashton, M. C., & Lee, K. (2001). A theoretical basis for the major dimensions of personality. *European Journal of Personality*, *15*(5), 327–353.

Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review*, *11*(2), 150–166.

Ashton, M. C., & Lee, K. (2008). The prediction of Honesty–Humility-related criteria by the HEXACO and Five-Factor Models of personality. *Journal of Research in Personality*, *42*(5), 1216–1228.

Ashton, M. C., Lee, K., & De Vries, R. E. (2014). The HEXACO Honesty–Humility, Agreeableness, and Emotionality factors: A review of research and theory. *Personality and Social Psychology Review*, *18*(2), 139–152.

Atwater, L. E., & Yammarino, F. J. (1992). Does self-other agreement on leadership perceptions moderate the validity of leadership and performance predictions? *Personnel Psychology*, *45*(1), 141–164.

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*(3), 335–344.

Becker, G. (2000). How important is transient error in estimating reliability? Going beyond simulation studies. *Psychological Methods*, *5*(3), 370–379.

Burisch, M. (1997). Test length and validity revisited. *European Journal of Personality*, *11*(4), 303–315.

Cattell, R. B. (1973). *Personality and mood by questionnaire*. San Francisco: Jossey-Bass.

Chmielewski, M., & Watson, D. (2009). What is being assessed and why it matters: The impact of transient error on trait research. *Journal of Personality and Social Psychology*, *97*(1), 186–202.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Costa, P. T., & McCrae, R. R. (1992). *NEO Personality Inventory—Revised (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) Professional Manual*. Odessa, FL: Psychological Assessment Resources.

De Fruyt, F., De Bolle, M., McCrae, R. R., Terracciano, A., & Costa, P. T. (2009). Assessing the universal structure of personality in early adolescence: The NEO-PI-R and NEO-PI-3 in 24 cultures. *Assessment*, *16*(3), 301–311.

De Vries, R. E. (2012). Personality predictors of leadership styles and the self–other agreement problem. *The Leadership Quarterly*, *23*(5), 809–821.

De Vries, R. E. (2013). The 24-item Brief HEXACO Inventory (BHI). *Journal of Research in Personality*, *47*(6), 871–880.

De Vries, R. E., Ashton, M. C., & Lee, K. (2009). De zes belangrijkste persoonlijkheidsdimensies en de HEXACO Persoonlijkheidsvragenlijst [The six most important personality dimensions and the HEXACO Personality Inventory]. *Gedrag & Organisatie*, *22*, 232–274.

De Vries, R. E., Lee, K., & Ashton, M. C. (2008). The Dutch HEXACO Personality Inventory: Psychometric properties, self–other agreement, and relations with psychopathy among low and high acquaintanceship dyads. *Journal of Personality Assessment*, *90*(2), 142–151.

De Vries, R. E., Realo, A., & Allik, J. (2016). *System and syntax files NEO and HEXACO personality item characteristics*. Retrieved from osf.io/uznpn

De Vries, R. E., Tybur, J. M., Pollet, T. V., & Van Vugt, M. (2016). Evolution, situational affordances, and the HEXACO model of personality. *Evolution & Human Behavior*, *37*, 407–421. doi:10.1016/j.evolhumbehav.2016.04.001.

De Vries, R. E., Wawoe, K. W., & Holtrop, D. (2016). What is engagement? Proactivity as the missing link in the HEXACO model of personality. *Journal of Personality*, *84*(2), 178–193. doi:10.1111/jopy.12150.

Deinzer, R., Steyer, R., Eid, M., Notz, P., Schwenkmezger, P., Ostendorf, F., & Neubauer, A. (1995). Situational effects in trait assessment: The FPI, NEOFFI, and EPI questionnaires. *European Journal of Personality*, *9*(1), 1–23.

Denissen, J. J., Geenen, R., Selfhout, M., & Van Aken, M. A. (2008). Single-item Big Five ratings in a social network design. *European Journal of Personality*, *22*(1), 37–54.

DiStefano, C., & Motl, R. W. (2006). Further investigating method effects associated with negatively worded items on self-report surveys. *Structural Equation Modeling*, *13*(3), 440–464.

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The Mini-IPIP scales: Tiny-yet-effective measures of the Big Five factors of personality. *Psychological Assessment*, *18*(2), 192–203.

Fleeson, W. (2001). Toward a structure-and process-integrated view of personality: Traits as density distributions of states. *Journal of Personality and Social Psychology*, *80*(6), 1011–1027.

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, *102*(4), 652–670.

Copyright © 2016 European Association of Personality Psychology

DOI: 10.1002/per

Funder, D. C., & Dobroth, K. M. (1987). Differences between traits: Properties associated with interjudge agreement. *Journal of Personality and Social Psychology*, *52*(2), 409–418.

Funder, D. C., Kolar, D. C., & Blackman, M. C. (1995). Agreement among judges of personality: interpersonal relations, similarity, and acquaintanceship. *Journal of Personality and Social Psychology*, *69*(4), 656–672.

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, *73*(2), 349–360.

Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, *59*(6), 1216–1229.

Gosling, S. D., Ko, S. J., Mannarelli, T., & Morris, M. E. (2002). A room with a cue: Personality judgments based on offices and bedrooms. *Journal of Personality and Social Psychology*, *82*(3), 379–398.

Greenberger, E., Chen, C., Dmitrieva, J., & Farruggia, S. P. (2003). Item-wording and the dimensionality of the Rosenberg Self-Esteem Scale: Do they matter? *Personality and Individual Differences*, *35*(6), 1241–1254.

Hendriks, A. A. J. (1997). *The construction of the Five-Factor Personality Inventory (FFPI)*. Groningen: Unpublished PhD Thesis, RijksUniversiteit Groningen.

Hendriks, A. A. J., Hofstee, W. K. B., & De Raad, B. (1999). The Five-Factor Personality Inventory. *Personality and Individual Differences*, *27*(2), 307–325.

Herzog, A. R., & Bachman, J. G. (1981). Effects of questionnaire length on response quality. *Public Opinion Quarterly*, *45*(4), 549–559.

Hofstee, W. K. B. (1991). *Richtlijnen voor het schrijven van vragenlijstitems [Directives voor writing questionnaire items]*. Groningen: Internal Note, Rijksuniversiteit Groningen.

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The Big Five Inventory-Versions 4a and 54*. Berkeley, CA: University of California, Berkeley, Institute of Personality and Social Research.

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*(4), 521–551.

Kallasmaa, T., Allik, J., Realo, A., & McCrae, R. R. (2000). The Estonian version of the NEO-PI-R: An examination of universal and culture-specific aspects of the Five-Factor Model. *European Journal of Personality*, *14*(3), 265–278.

Kraut, A. I., Wolfson, A. D., & Rothenberg, A. (1975). Some effects of position on opinion survey items. *Journal of Applied Psychology*, *60*(6), 774–776.

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO personality inventory. *Multivariate Behavioral Research*, *39*(2), 329–358.

Leitsalu, L., Haller, T., Esko, T., Tammesoo, M.-L., Alavere, H., Snieder, H., . . . Milani, L. (2014). Cohort profile: Estonian biobank of the Estonian Genome center, University of Tartu. *International Journal of Epidemiology*, *44*(4), 1137-1147.

Marsh, H. W. (1986). Negative item bias in ratings scales for preadolescent children: A cognitive-developmental phenomenon. *Developmental Psychology*, *22*(1), 37–49.

McCrae, R. R. (2015). A more nuanced view of reliability: Specificity in the trait hierarchy. *Personality and Social Psychology Review*, *19*(2), 97–112.

McCrae, R. R., Costa, P. T., & Martin, T. A. (2005). The NEO–PI–3: A more readable revised NEO personality inventory. *Journal of Personality Assessment*, *84*(3), 261–270.

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review*, *15*(1), 28–50.

Meng, X. l., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin*, *111*(1), 172–175.

Mõttus, R., Allik, J., Hřebíčková, M., Kööts-Ausmees, L., & Realo, A. (2016). Age differences in the variance of personality characteristics. *European Journal of Personality*, *30*, 4–11. doi:10.1002/per.2036.

Mõttus, R., McCrae, R. R., Allik, J., & Realo, A. (2014). Cross-rater agreement on common and specific variance of personality scales and items. *Journal of Research in Personality*, *52*, 47–54.

Mõttus, R., Pullmann, H., & Allik, J. (2006). Toward more readable Big Five personality inventories. *European Journal of Psychological Assessment*, *22*(3), 149–157.

Ostroff, C., Atwater, L. E., & Feinberg, B. J. (2004). Understanding self–other agreement: A look at rater and ratee characteristics, context, and outcomes. *Personnel Psychology*, *57*(2), 333–375.

Quirk, S. W., Christiansen, N. D., Wagner, S. H., & McNulty, J. L. (2003). On the usefulness of measures of normal personality for clinical assessment: Evidence of the incremental validity of the Revised NEO Personality Inventory. *Psychological Assessment*, *15*(3), 311–325.

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, *41*(1), 203–212.

Rauthmann, J. F. (2012). You say the party is dull, I say it is lively: A componential approach to how situations are perceived to disentangle perceiver, situation, and perceiver × situation variance. *Social Psychological and Personality Science*, *3*(5), 519–528.

Realo, A., Teras, A., Kööts-Ausmees, L., Esko, T., Metspalu, A., & Allik, J. (2015). The relationship between the Five-Factor Model personality traits and peptic ulcer disease in a large population-based adult sample. *Scandinavian Journal of Psychology*, *56*(6), 693–699.

Saucier, G., & Golberg, L. R. (2002). Assessing the Big Five: Applications of 10 psychometric criteria to the development of marker scales. In B. De Raad, & M. Perugini (Eds.), *Big Five assessment* (pp. 30–54). Ashland, OH, US: Hogrefe & Huber Publishers.

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual-differences constructs. *Psychological Methods*, *8*(2), 206–224.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, *44*(6), 922–932.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107–120.

Spörrle, M., & Bekk, M. (2014). Meta-analytic guidelines for evaluating single-item reliabilities of personality instruments. *Assessment*, *21*(3), 272–282.

Suzuki, T., Samuel, D. B., Pahlen, S., & Krueger, R. F. (2015). DSM-5 alternative personality disorder model traits as maladaptive extreme variants of the five-factor model: An item-response theory analysis. *Journal of Abnormal Psychology*, *124*(2), 343–354.

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology*, *88*(3), 500–517.

Thalmayer, A. G., Saucier, G., & Eigenhuis, A. (2011). Comparative validity of brief to medium-length Big Five and Big Six Personality Questionnaires. *Psychological Assessment*, *23*(4), 995–1009.

Thorndike, R. L. (1951). Reliability. In E. F. Linquist (Ed.), *Educational measurement* (pp. 560–620). Washington D.C.: American Council on Education.

Wanous, J. P., & Hudy, M. J. (2001). Single-item reliability: A replication and extension. *Organizational Research Methods*, *4*(4), 361–375.

Wanous, J. P., & Reichers, A. E. (1996). Estimating the reliability of a single-item measure. *Psychological Reports*, *78*(2), 631–634.

Warr, P., & Bourne, A. (1999). Factors influencing two types of congruence in multirater judgments. *Human Performance*, *12*(3–4), 183–210.

Copyright © 2016 European Association of Personality Psychology

DOI: 10.1002/per

Watson, D., Hubbard, B., & Wiese, D. (2000). Self–other agreement in personality and affectivity: The role of acquaintanceship, trait visibility, and assumed similarity. *Journal of Personality and Social Psychology*, *78*(3), 546–558.

Whittington, J. L., Coker, R. H., Goodwin, V. L., Ickes, W., & Murray, B. (2009). Transactional leadership revisited: Self–other agreement and its consequences. *Journal of Applied Social Psychology*, *39*(8), 1860–1886.

Wiggins, J. S. (2003). *Paradigms of personality assessment*. New York: Guilford Press.

Wood, D., & Wortman, J. (2012). Trait means and desirabilities as artifactual and real sources of differential stability of personality traits. *Journal of Personality*, *80*(3), 665–701.

Zettler, I., Lang, J. W., Hülsheger, U. R., & Hilbig, B. E. (2016). Dissociating indifferent, directional, and extreme responding in personality data: Applying the Three-Process Model to self-and observer reports. *Journal of Personality*, *84*(4), 461–472. doi:10.1111/jopy.12172.

Ziegler, M., Poropat, A., & Mell, J. (2014). Does the length of a questionnaire matter? *Journal of Individual Differences*, *35*(4), 250–261.

Copyright © 2016 European Association of Personality Psychology

DOI: 10.1002/per

Copyright of European Journal of Personality is the property of John Wiley & Sons, Inc. and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.