# Assessing Meaningful Within-Person Variability in Likert-Scale Rated Personality Descriptions: An IRT Tree Approach

Jonas W. B. Lang
Ghent University

Filip Lievens
Singapore Management University

Filip De Fruyt
Ghent University

Ingo Zettler
University of Copenhagen

Jennifer L. Tackett
Northwestern University

**in press, Psychological Assessment**



Author Note

Jonas W. B. Lang, Department of Personnel Management and Work and Organizational Psychology, Ghent University, Belgium; Filip Lievens, Lee Kong Chian School of Business, Singapore Management University, Singapore; Filip De Fruyt, Department of Developmental, Personality and Social psychology, Ghent University, Belgium; Ingo Zettler, Department of Psychology, University of Copenhagen; Jennifer L. Tackett, Department of Psychology, Northwestern University.

Study 2 was supported by a grant of the Fund for Scientific Research- Flanders (G092512N) awarded to Filip Lievens.

Correspondence concerning this article should be addressed to Jonas W. B. Lang, Henri Dunantlaan 2, 9000 Ghent, Belgium, E-mail: jonas.lang@ugent.be.

## Abstract

Personality researchers and clinical psychologists have long been interested in within-person variability in a given personality trait. Two critical methodological challenges that stymie current research on within-person variability are separating meaningful within-person variability from 1) true differences in trait level and 2) careless responding (or person unreliability). To partly avoid these issues, personality researchers commonly only study within-person variability in personality states over time using the standard deviation (*SD*) across repeated measurements of the same items (typically across days)—a relatively resource-intensive approach. In this article, we detail an approach that allows researchers to measure another type of within-person variability. The described approach utilizes IRT on the basis of Böckenholt's (2012) three-process model, and extracts a meaningful variability score from Likert-ratings of personality descriptions that is distinct from directional (trait) responding. Two studies ($N = 577$; $N = 120$-235) suggest that IRT variability generalizes across traits, has high split-half reliability, is not highly correlated with established indices of IRT person unreliability for directional trait responding, and correlates with within-person *SD*s from personality inventories and within-person *SD*s in a diary study with repeated measurements across days 20 months later. The implications and usefulness of IRT variability from personality descriptions as a conceptually clarified, efficient, and feasible assessment of within-person variability in personality ratings are discussed.


Keywords: Variability, flux, adaptability, psychometrics, tree models

## Public Significance Statements

An innovative psychometric method allows researchers to measure the degree to which persons show variability in their personality. This new personality variability score can be estimated on the basis of common Likert-based personality questionnaires

# Assessing Meaningful Within-Person Variability in Likert-Scale Rated Personality Descriptions: An IRT Tree Approach

Within-person variability in a given personality has been described as oscillation (Flügel, 1929; Spearman & Jones, 1950), steadiness of character (Walton, 1936), personality flexibility (Paulhus & Martin, 1988), or flux (Moskowitz & Zuroff, 2004), and has been linked to various outcomes including adaptability and strength (Fiske & Rice, 1955; Pulakos, Arad, Donovan, & Plamondon, 2000), or, conversely, emotional dysregulation (e.g., Allport's *Letters from Jenny*, 1965). There are two different ways to conceptualize within-person variability (Fiske & Rice, 1955; Lievens et al., 2017). One conceptualization focuses on within-person variability as variability in personality states across time. To measure this type of variability, personality researchers typically use the within-person *SD* across repeated measurements (typically across days) of the same items that assess a particular personality state (Baird, Le, & Lucas, 2006; Eid & Diener, 1999; Fleeson, 2001). A limitation of this variability-across-time approach is that it is resource-intensive because participants need to participate in time-consuming studies, e.g., longer diary or experience sampling studies. However, an important practical advantage of studying variability across time is that it partly avoids a set of methodological issues.

A second conceptualization of within-person variability is to define it as intra-individual variability across the personality descriptions in personality trait measures. Personality trait measures commonly ask respondents to rate to what degree a set of statements describing habits, feelings, preferences or behaviors in common situations is typical for them (Werner & Pervin, 1986) using Likert scales. A potential practical advantage of focusing on variability-across-descriptions is that it does not require a design with repeated measurements across days. However, the variability-across-descriptions conceptualization is methodologically challenging because variability in responses to different statements can also reflect other characteristics of the

measurement process (Eid & Diener, 1999; Fiske & Rice, 1955). Most notably, 1) the true trait level and 2) the distribution of the item difficulties across the trait continuum both have the potential to affect the most common measure of within-person variability—the within-person *SD*. Furthermore, it is also possible that observed variability in responses actually reflects aberrant or careless responding (also described as person misfit or person unreliability in the item-response theory literature; Meijer & Sijtsma, 2001; Reise, 1995; Snijders, 2001).

In this article, we focus on the variability-across-descriptions conceptualization and detail how item-response theory (IRT) can be used to extract meaningful variability across personality descriptions from Likert-scale personality inventory data. The goals of our article were to (a) address some conceptual challenges in the within-person variability literature from an IRT perspective, and (b) to complement established research strategies for capturing within-person variability across time with a research approach that can be used to capture within-person variability across personality descriptions from personality inventories and can therefore readily be applied to most personality datasets. The described approach builds on Böckenholt's (2012) three-process IRT tree model for Likert-scale data and earlier applications of this model in personality research (Zettler, Lang, Hülsheger, & Hilbig, 2016). Recently, researchers have suggested that this model can be modified to separate latent IRT scores for directional (trait) responding from IRT variability scores (Lang, Tackett, & Zettler, 2017; Lievens, 2017; Lievens et al., 2017). The underlying rationale for IRT variability scores is to summarize meaningful variability that is distinct from measurement-specific sources of variability (i.e., caused by other characteristics of the measurement process). This article is the first study of which we are aware that uses this approach to extract meaningful within-person variability from personality trait measures, and thus examines the broad feasibility of this approach for personality and clinical research.

## Within-Person Variability in Personality

Personality researchers have long suggested that personality traits typically show variability within persons that goes beyond measurement error. Personality and clinical psychologists frequently observed and discussed variability in diary or longitudinal observations of single patients. For instance, Gordon Allport's study of an anonymous woman named Jenny showed considerable levels of variability associated with her emotional dysregulation (Allport, 1965), and, in particular, emphasized the importance of considering substantive variability as an independent and incremental construct, above and beyond average trait levels. Early empirical studies of individual differences in variability originally started in experimental psychology and focused on variability across laboratory settings (Flügel, 1929; Hollingworth, 1925; Kehr, 1916). However, this type of work quickly also influenced personality psychologists, who suggested that the absence of variability could be interpreted as a person's steadiness of the character (Walton, 1936). Building on this idea, Charles Spearman (e.g., Spearman & Jones, 1950) suggested that an important variability factor exists that is independent from his g factor of intelligence. He referred to this factor as "oscillation".

Interest in the study of within-person variability among personality psychologists and clinical researchers may also originate from the fact that the notion of variability or consistency is tied to the trait concept (Mischel, 1968; Pervin, 1994; Winter, John, Stewart, Klohnen, & Duncan, 1998). Gordon Allport (1937) observed that "the existence of a trait always comes from a demonstration by some acceptable method of *consistency* in behavior" (p. 330). In a similar vein, Hans Eysenck has emphasized that the notion of correlation is central for the trait concept (Eysenck, 1953; Eysenck & Eysenck, 1985). In line with this definition, a common method to establish traits are correlational studies and factor analyses of personality descriptions with the goal to identify individuals' "typical" behavioral conduct. Personality psychologists have also

suggested that the acceptance of a personality trait measures commonly requires “tests of internal and cross-situational consistency, as well as temporal stability” (Winter et al., 1998). However, Allport (1937) also suggested that perfect consistency should not be expected because “traits are often aroused in one situation but not in another; not all stimuli are equivalent in effectiveness” (pp. 331-332). A logical next step from Allport’s observations is that individuals may systematically differ in the degree to which they show consistency (Fiske & Rice, 1955; Fleeson, 2001; Fleeson & Jayawickreme, 2015). Research on within-person variability typically interprets variability as a distinct characteristic of personality beyond average trait levels.

From a conceptual perspective, within-person variability can be construed in two different ways (Fiske & Rice, 1955; Lievens et al., 2017). First, within-person variability can refer to variability across time in a particular personality state. Researchers have suggested that this type of intra-individual variability in personality states captures changes (increases or decreases) in individuals’ action tendencies or frequency distributions of behavior (Conner, Tennen, Fleeson, & Barrett, 2009; Fleeson, 2001; Fleeson & Jayawickreme, 2015). Although typically described as variability across time, it is important to be aware that this type of variability may in large part be driven by the different types of situations the individual experiences in his or her daily life and how he/she expresses (or describes) his/her personality in these situations.

A second way to conceptualize intra-individual variability is to define it as intra-individual variability across the personality descriptions in a personality trait measure. Personality trait measures commonly describe a range of habits, typical feelings, preferences or behaviors in common situations entailed in the same personality trait (Werner & Pervin, 1986). This alternative variability-across-descriptions conceptualization of within-person variability focuses on the degree to which a trait is being differentially activated, suppressed, or expressed across a range of habits, preferences or behaviors that belong to the same class of typical

behavior, and is thus more related to the traditional method to demonstrate consistency through correlational or factor analyses of personality descriptions.

## Extant Approaches for Studying Within-Person Variability in Personality

Contemporary research on within-person variability largely focuses on the variability across time perspective and uses either diary or experience sampling designs (ESM; Conner et al., 2009; Eid & Diener, 1999; Fleeson, 2001). In these studies, respondents fill out the same short questionnaire (or parallel versions) across several measurement occasions (typically days or several measurements within days). Meaningful within-person variability is then operationalized using the *SD* across several measurement occasions (Conner et al., 2009; Eid & Diener, 1999; Fleeson, 2001). Table 1 provides a fictitious example of the assessment of within-person variability across days using this approach. Commonly, within-person *SD*s across time for different personality states are substantially correlated (Baird et al., 2006; Eid & Diener, 1999; Fleeson, 2001) and correlations are frequently in the high .30 to .60s and are thus as high as between facets within broad traits (Costa & McCrae, 1992, 1995; Soto & John, 2009). Researchers have therefore suggested that a single variability trait—consistent with Spearman's idea of a general oscillation factor—may exist (Baird et al., 2006) in addition to dimension-specific variability traits. One major advantage of the diary/ESM approach is that it partly addresses some methodological challenges. Specifically, using the same measurement scale with multiple items and comparing persons with themselves across days partly eliminates non-meaningful variability when one assumes that the non-meaningful variability exists in each measurement of a person to the same degree. However, a core disadvantage of focusing on variability-across-time is that the measurement of this type of within-person variability is resource intensive because respondents need to carry around their diaries or ESM devices. Respondents also need to be convinced to participate in an assessment procedure for several days

such that a representative sample of days or timepoints for the individual person can be gathered. To our knowledge, the approach is therefore less frequently used in clinical and other assessment practice than other forms of assessment (e.g., single time point questionnaires). There may also be many assessment applications in which the use of this type of approach is not feasible like, for instance, when a client/patient visits for just a single assessment session. One alternative to the ESM/diary approach is to directly ask respondents to retrospectively report variability over time (Bem & Allen, 1974; Chaplin & Goldberg, 1985; Fleisher, Woehr, Edwards, & Cullen, 2011; Paulhus & Martin, 1988). For instance, researchers have asked participants to report to what percentage out of a total of 100 percentage points the item "I am the life of the party" was "very inaccurate", "neither inaccurate nor accurate", and "very accurate" in the last six months (Fleisher et al., 2011).

In this article, we contribute to the literature on within-person variability by describing another approach. We focus on within-person variability across Likert-scale rated personality descriptions as another conceptualization of within-person variability instead of within-person variability over time. Researchers have long been aware of the possibility to gather information on within-person variability from common personality measures. However, we are not aware of much research that has focused on this type of within-person variability. A potential explanation for this lack of research is that researchers who are interested in within-person variability across personality descriptions face several methodological challenges. To address these challenges, we use an IRT tree model to capture meaningful within-person variability from Likert-scale rated personality descriptions. The core goal of our paper is less the invention of a new statistical method—the IRT model we use is a variant of a common type of an IRT tree model that can be estimated as a basic generalized linear mixed-effect model. Rather, we seek to show how an elegant psychometric model can be used to solve an important methodological challenge in this

research field. The aim of our paper is also not to criticize earlier work on within-person variability. Instead, we seek to broaden this work by developing new ways to make the construct more readily available to researchers.

## Challenges in Studying Within-Person Variability in Personality Descriptions

Although personality psychologists had long been interested in variability constructs, there are many inherent challenges in studying within-person variability (Baird et al., 2006; Eid & Diener, 1999; Fiske & Rice, 1955; Meijer & Sijtsma, 2001; Schmitt, Chan, Sacco, McFarland, & Jennings, 1999). In addition to substantive interpretations of within-person variability as flexibility, adaptability, trait steadiness or emotion regulation, there are also alternative sources for variability in a response pattern to a set of personality items. In particular, researchers face three challenges when they seek to separate meaningful within-person variability from other sources of variability in responses (e.g., measurement-specific variability).

### Meaningful Variability vs. Measurement Error

The first challenge is to separate meaningful within-person variability from variability because of measurement error. One way to describe this problem from a psychometric perspective is to imagine a test that follows the 1-parameter IRT or Rasch model and includes items that describe a range of behaviors or habits that belong to the same personality trait. The problem can better be illustrated using IRT because classical test theory makes unrealistic assumptions (constant *SE* across the trait continuum) on the nature of measurement error (Hambleton, Swaminathan, & Rogers, 1991). In the context of IRT, both items and persons are placed on a continuum. In a 1PL or Rasch model, each item *j* has a difficulty parameter $\gamma_j$ (Greek letter gamma) and each person *k* receives a latent trait score $\theta_k$ (Greek letter theta). Depending on how the item difficulty parameters $\gamma_j$ are distributed across the continuum, the measurement precision of individual test scores also varies. For instance, when a test mostly contains items

with item difficulties ($\gamma_j$) at the bottom of the latent trait distribution, persons with latent trait scores ($\theta_k$) close to the item difficulties should show relatively more variability in their observed behavior and smaller measurement errors ($SE_{\theta k}$) than persons with latent trait scores at the top of the distribution. The reason is that a Rasch/1PL test provides more information or measurement precision when most item difficulties are close to the latent trait score $\theta_k$ of a person. As another example, take the response patterns to a test with all item difficulties at $\gamma_j = 0$ of two persons with $\theta_1 = 0$ (middle of the distribution) and $\theta_2 = 3$ (top of the distribution). The response pattern for person 1 with $\theta_1 = 0$ would show maximum variability (50% 1 and 50% 0 responses when the response is dichotomous, as in a typical Rasch or 1PL model) but also small measurement error. In contrast, the response pattern of person 2 with $\theta_2 = 3$ would almost exclusively show 1 responses and only occasionally a 0 response in-between the 1 responses. Person 2 would thus show low variability and high measurement error. These examples show that within-person variability in a response pattern and the associated measurement error is entangled with both the item difficulty parameters of a test ($\gamma_j$) and the latent trait estimates ($\theta_k$) of the persons.[1] Depending on the item difficulties of a specific test, a person with a particular $\theta_k$ may have high or low variability and variability generally varies across the trait continuum because of measurement error.

**Meaningful Variability vs. Ceiling Effects**

The second challenge for research on within-person variability is how to deal with potential ceiling effects. The commonly used measure of within-person variability is the within-person standard deviation (*SD*). A potential problem with the within-person *SD* is that it is functionally dependent on the trait level because *SD* asymptotes to zero at the highest or lowest

[1]In more complex IRT models like the 2PL, measurement error is additionally influenced by the item discrimination parameters.

scores on a personality trait. From an IRT perspective, the (floor or) ceiling problem is an extreme case of the measurement error problem we described as the first challenge for within-person research. IRT suggests that the observed *SD* should be zero or very close to zero when the item difficulties $\gamma_j$ are in the middle or the bottom of the distribution and a person has an extremely high trait level (or $\theta_k$ in the context of IRT). In this scenario, the observed response pattern shows minimal or no variability (and, thus measurement error is high). Likewise, IRT would also suggest that the observed *SD* should be zero or close to zero in the opposite scenario (high or average item difficulties and extremely low trait level $\theta_k$).

However, the fact that the *SD* asymptotes to zero at high and low trait levels is not only a psychometric problem but also a conceptual problem when researchers seek to extract meaningful within-person variability scores from personality descriptions. Persons cannot simultaneously be (measured as) high on the trait and (measured as) high on variability; thus, the trait direction and variability on the trait are not fully psychometrically distinct. Table 2 illustrates this problem using the hypothetical responses of three persons to four personality items with a 5-point Likert scale from 1 (strongly disagree) to 5 (strongly agree). As shown in Table 2, Person 1 and Person 3 both have identical mean scores on the trait (2.5). However, Person 3 shows more variability. In contrast, Person 2 has a higher trait score (5). His or her variability is 0 and cannot be higher unless his or her trait score would decrease. Mean level and variability are accordingly functionally entangled.

**Meaningful Variability vs. Person Misfit**

The third challenge for within-person variability research is the need to distinguish between within-person variability and the IRT construct of person fit or person unreliability (Magis, Raiche, & Beland, 2012; Meijer & Sijtsma, 2001; Meijer & Tendeiro, 2012; Schmitt et al., 1999; Snijders, 2001). Person fit indices are commonly interpreted as indicators of careless

responding, low-attention responding, or even cheating. The goal is typically to detect aberrant response patterns in datasets. Most person fit-indices are directly based on IRT and are designed to detect respondents who deviate from a plausible response pattern. For instance, a respondent who answers many difficult items correctly but many easy items incorrectly in an ability test would be flagged. Correspondingly, in the context of a personality test, a person with a high latent trait score $\theta_k$ who unexpectedly disagrees with some "easy" items (low item difficulty $\gamma_j$) would also receive a high person misfit estimate. The most advanced and commonly recommended (Magis et al., 2012; Meijer & Tendeiro, 2012) person-fit index is Snijders's $l_z^*$ (Snijders, 2001). This person-fit index accounts for the latent trait scores of the persons $\theta_j$ and the item difficulties. Because person-fit indices are designed to index aberrations from an IRT model, these indices are not directly incorporated in or a part of the IRT model. Researchers have shown that an earlier version of $I_z$* is related to test-taking motivation and conscientiousness and that eliminating respondents with high person misfit scores typically improves test validity (Schmitt et al., 1999). These findings are generally in line with the common interpretation of person fit indices as indices of aberrant responding.

## An IRT Tree Approach to Meaningful Within-Person Variability

Tree models (or nested models) are a type of choice model that has long been used in economics (Greene, 2012), for instance, in Nobel prize winner Daniel McFadden's work (McFadden, 2001). Tree models predict the average choices people make in response to a series of decision problems. IRT tree models (Böckenholt, 2012; De Boeck & Partchev, 2012; Jeon & De Boeck, 2016) extend tree models by allowing for systematic individual differences in choices between persons. The core idea behind tree models is that decisions between several alternatives can be split up into several subdecisions that are commonly called pseudoitems for IRT tree models. In IRT tree models, the pseudoitems vary both across the items of the test and persons

also vary in the degree to which they react to particular pseudoitems. IRT tree models typically estimate a separate latent trait for each pseudoitem. One tree model that is of particular interest for personality research is the three-process model suggested by Böckenholt (2012). This model was developed for Likert-scale data and splits the information from Likert scales into the direction of the trait and two components that are functionally independent from the direction of the trait and capture how strongly the trait is expressed in behavior. The model achieves this through a pseudoitem that differentiates between midpoint responding (or response refusal), a pseudoitem that captures direction responding, and finally a pseudoitem that differentiates between modest and extreme responding (see Figure 1A). An initial study with personality trait data (Zettler et al., 2016) revealed that direction responding is extremely highly correlated with the full scale scores, and thus is highly similar to the full scale scores. The two other pseudoitems (midpoint responding and extreme responding) are typically only weakly correlated with direction responding but are typically moderately to highly correlated with each other (correlations are provided in the results section). Lievens et al. (2017) observed similar patterns in a situational judgement test using a Likert-type scale. The fact that the three-process model functionally splits data from Likert scales into a directional or trait component (individual differences in the direction pseudoitem) and two components that capture response variability (midpoint responding and extreme responding) makes the approach interesting for studying meaningful within-person variability (Lang et al., 2017; Lievens, 2017; Lievens et al., 2017). Building on these ideas, we apply a modified version of the Böckenholt model with a uniform variability-related latent trait (see Figure 1B) to personality data in this article. We refer to the model as the trait variability tree model (TVTM) and we label the variability latent trait in the TVTM as IRT variability. IRT variability captures a tendency of respondents to endorse response options that are extremes on the scale (high or low) and functionally independent of the direction

of the trait itself. This approach has a couple of potential advantages.

First, IRT variability can clearly be distinguished from person fit in the context of IRT models. From an IRT perspective, only variability due to measurement error (Hambleton et al., 1991) and due to person misfit (Meijer & Tendeiro, 2012) has been described in earlier research. The TVTM complements these two types of IRT variability with a meaningful type of variability in item responses.

Second, IRT variability provides a definition of variability at the level of the individual item. In contrast, the within-person *SD* in other research is commonly only defined with respect to a set of items or repeated measurements of the same item. The slightly different definition of the variability latent trait has direct implication for the maximum score problem commonly encountered in within-person variability research. Specifically, a maximum latent trait score is possible in combination with a high level of IRT variability. The reason is that the latent trait scores are based on pseudoitem II while IRT variability is based on pseudoitems I and III.

Finally and third, IRT variability is functionally distinct from the item difficulties and from the IRT trait score (directional responding) because the IRT model accounts for both sources of variability in response patterns. As a result, the approach can be used on items with different content (it is not necessary to have repeated measurements of the same items) and can directly be applied to content-diverse Likert-scale rated personality descriptions.

## Empirical Studies

Our overarching goal for this article was to describe and study an IRT approach that makes the extraction of meaningful within-person variability from common personality trait measures feasible for researchers. We therefore supplement our theoretical description of the trait variability tree model (TVTM) and the IRT variability and IRT trait scores from this model by studying the empirical characteristics and usefulness of these scores in two datasets. In both

datasets, we examine the split-half reliability of IRT variability scores, the degree to which IRT variability generalizes across traits, and the degree to which IRT variability is correlated with an established index of IRT person misfit for directional trait responding. In the second dataset, we additionally explore whether IRT variability across personality descriptions in a common personality inventory as one conceptualization of within-person variability is linked to the more commonly used alternative conceptualization of within-person variability as variability in personality states across time (operationalized as the within-person *SD*s across repeated measurements in a diary design). As we detailed, the underlying conceptualizations of both types of within-person variability markedly differ both theoretically and conceptually. In our study, the two measurements were also considerably separated in time (20 months) so that our study is a strong test of the idea that the assessment of within-person variability across personality descriptions using the TVTM IRT model can predict outside criteria.

## Method

### Study 1: Self- and Other Reports

The first dataset we examined has originally been studied by Zettler et al. (2016). It includes a total of 577 self- and observer reports on personality. The observers all were psychology students with a mean age of 21 years (*SD* = 3, range = 18–54) who provided the observer reports as part of an undergraduate personality psychology course. The self-reporters were age diverse with a mean age of 30 years (*SD* = 14, range = 16–72) and provided informed consent. Fifty-one percent of the target participants and 73% of the observers were female. The ethical committee of the psychology faculty of Maastricht University approved the study.

Both self-ratings and observers ratings of personality were provided using the long version of the HEXACO Personality Inventory-Revised (Lee & Ashton, 2006)—a broadband personality inventory that captures six major personality dimensions (Honesty-Humility,

Emotionality, eXtraversion, Agreeableness, Conscientiousness, and Openness to Experience). The underlying HEXACO model is similar to but also differs in some regards from the more well-known five-factor model of personality (Lee & Ashton, 2008). The six dimensions are captured using a total of 192 items (each dimension is assessed using 32 items) and are rated on a Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree).

**Study 2: Self-Reports and Diary Study**

The second study included a total of 235 industrial and organizational psychology students (77 percent female; average age = 23.2 years) who initially completed a series of psychometric measures for course credit and a detailed feedback report (Time 1, T1) and has earlier been studied by Lievens et al. (2018). They were contacted again about 20 months later. Out of the 235 initial participants, 120 agreed to participate in a diary study in exchange for a gift voucher over 30 Euros (Time 2; T2). Data were collected in accordance with the ethical guidelines for research involving human subjects by Ghent University, Belgium (G092512N).[2]

The self-report personality inventory at T1 included the traits of the five-factor model of personality (Extraversion, Conscientiousness, Agreeableness, Neuroticism, and Openness to Experience). For each of the five personality traits, the inventory included ten items from the International Personality Item Pool (Goldberg, 1999). The items were rated on a scale ranging from 1 (very inaccurate) to 7 (very accurate). For the purpose of this study, we collapsed the two non-extreme endorsement categories (somewhat and slightly accurate/inaccurate) into one category (accurate/inaccurate) to simplify the analysis on the basis of the trait variability tree model (TVTM).[3]

[2] The earlier manuscript focused on an situational judgment test while we focus on the personality inventory which has not been considered in the earlier work.

[3] Collapsing categories is generally an accepted approach to make IRT analyses more parsimonious (Doran

The diary study at T2 lasted ten weekdays and participants filled in a personality inventory on their computer or mobile device at the evening of each day. The inventory included a total of 22 adjectives measuring Agreeableness, Extraversion, and Conscientiousness, and the instructions asked participants to describe themselves during the day (and not in general) using a Likert scale, ranging from 1 (completely not characteristic) to 5 (very characteristic). Internal consistency reliability estimates for the means across days ranged from .63 to .83.

**Analytical Strategy**

The trait variability tree model (TVTM) can formally be written in several equations (see Table 3). Each equation provides the probability $\gamma_{jk}$ that a particular person *k* provides a particular type of response for a particular item *j*. Tree models split the probability of a response into several pseudoitems that capture the decisions in the response tree. The TVTM includes three pseudoitems. The first pseudoitem I captures midpoint responding (coded 1 = midpoint not endorsed vs. 0 = midpoint endorsed). The second pseudoitem II captures directional or trait responding (coded 1 = agree vs. 0 = disagree). Finally, the third pseudoitem II captures extreme responses coded 0 = extreme response vs. 1 = no extreme response. The equations describe $\gamma_{ij}$ as a function of the item difficulties $\gamma_j^{(I)}$, $\gamma_j^{(I)}$, and $\gamma_j^{(I)}$ for the three pseudoitems I, II, and II, respectively. The model also includes one or more latent traits for pseudoitem II (trait responding). In the equation below, dimension-specific estimates, $\theta_k^{(IIA)}$, $\theta_k^{(IIB)}$, ... $\theta_k^{(IIX)}$ are for several latent direction traits A, B, … X and it depends on the content of the item which of these traits is active. Finally, the model includes a latent trait that captures individual differences

et al., 2007; Stark, Chernyshenko, Drasgow, & Williams, 2006). However, we also re-ran the analyses in this paper using a model with four pseudoitems. The analyses were highly similar and the conclusions did not change. We therefore report the more simple analyses.

in pseudoitem I and III (trait variability), $\theta_k^{(V)}$. A variant of the model is to specify content-dimension specific variability traits, $\theta_k^{(VA)}$, $\theta_k^{(VB)}$, ... $\theta_k^{(VX)}$. Note that in all equations, Φ refers to the cumulative standard normal distribution.

Although the model specification may appear relatively complex, the model is in fact just a combination of three Rasch/1PL models ($\gamma_j + \theta_k$) that are nested in each other and each of the equations is a product of the item and person parameters from each of these models. A simple example for how the model works would be a hypothetical person 197 with a variability latent trait $\theta_{k=197}^{(V)} = 1$ and an agreeableness direction trait $\theta_{k=197}^{(IIB)} = 0$ that works on a hypothetical agreeableness item 23 with item difficulties $\gamma_{j=23}^{(I)} = 1$, $\gamma_{j=23}^{(II)} = 0.5$, and $\gamma_{j=23}^{(III)} = 0.25$ for for Pseudoitem I, Pseudoitem II, and Pseudoitem III, respectively. The probability that this person would answer with agree would be

$$\Pr(\gamma_{j=23;\,k=197} = \text{agree}) = \Phi\left(\gamma_{j=23}^{(I)} + \theta_{k=197}^{(V)}\right)\Phi\left(\gamma_{j=23}^{(II)} + \theta_{k=197}^{(IIB)}\right)\Phi\left(\gamma_{j=23}^{(III)} - \theta_{k=197}^{(VE)}\right) = \Phi(1+1)\,\Phi(0.5+0)\Phi(0.25-1) \approx 0.153.$$

Like in other IRT models, item-level variability in the model can be modeled as fixed effects (i.e., persons as random, items as fixed) like in Böckenholt's (2012) article or alternatively as random effects (De Boeck, 2008; De Boeck & Partchev, 2012; Zettler et al., 2016). For the purpose of simplicity, we report the models for random item-effects throughout this paper. Especially for personality items, the assumption of random effects (items are sampled from a larger universe of items) makes particular sense (Zettler et al., 2016). The latent traits in the model are commonly allowed to correlate with covariance matrix $\Sigma_1$.

The described model can be estimated in most generalized linear mixed-effect modeling (glmm) software packages and also in some structural equation modeling software (e.g., Böckenholt, 2012). When tree models are analyzed using glmm software, a long format is used that allows the number of item responses to vary across persons. This approach is useful for tree models because the number of responses can vary depending on the nature of the response tree. For instance, in the TVTM, a person that endorses the mid-category only has one response for the item while all other responses yield information on three pseudoitems. In the present study, we used the freely available lme4 package (Bates, Maechler, Bolker, & Walker, 2015) for R (R Core Team, 2014) that is frequently used for IRT analyses (De Boeck, 2008; De Boeck & Partchev, 2012; Doran, Bates, Bliese, & Dowling, 2007). For data preparation, we relied on the irtrees package (De Boeck & Partchev, 2012).

## Results

### Model Evaluation and Split-Half Reliabilities

We started our analyses by fitting the TVTM with an overall variability trait to the self-report personality inventories in both studies. The model estimates for Study 1 and 2 are provided in Table 4 and 5, respectively. As indicated by the variance estimates in Tables 4 and 5, there were considerable individual differences between persons in both the direction (or content) latent traits as well as the IRT variability trait. IRT variability was also clearly distinct from the content-traits (low correlations between -.17 and .23). One critical question is the degree to which overall IRT variability can reliably be measured and thus generalizes from one set of personality items to another set of personality items. To study this question, we split the self-report personality inventories in both studies into two halves and re-estimated the TVTM models. The first half was the items presented first in the personality inventories. The results are shown in Tables 6 and 7 and revealed that the split-half reliabilities between IRT variability in the two halves were $r^*_{tt}$ =

.92 and $r^*_{tt}$ = .83 for Study 1 and 2, respectively.

To further evaluate the fit of the model, we examined two alternative models. The first model was the full three-process model with individual differences in each pseudoitem and for each content (trait) dimension. In this model, IRT variability splits into latent traits for Pseudoitem I and Pseudoitem III for each dimension. The correlation between these traits ranged from *r* = -.37 to *r* = -.56 (average *r* = -.45) in Study 1, and from *r* = -.34 to *r* = -.87 (average *r* = -.64) in Study 2. Furthermore, the two pseudoitems were also highly related at the item level (*r* = .66 in Study 1 and *r* = .80 in Study 2; also see Tables 4 and 5). The fact that indifference and extremity were negatively correlated at both the person and the item level suggests that both dimensions capture functionally related types of variability. The correlations at the person level are in the range or exceed commonly observed correlations between facet scales in personality inventories that are typically between .30 and .60 (Costa & McCrae, 1992, 1995; Soto & John, 2009).[4] The findings thus provide some support for using a single variability dimension which is consistent with the existing literature on within-person variability.

The second alternative model included separate IRT variability dimensions for each content dimension. Results revealed that the IRT variability dimensions in the longer 192 item personality inventory used in Study 1 were strongly correlated with each other (average *r* = .61; range: *r* = .49 to *r* = .78) and clearly distinct from the content dimensions (see Table 4). We also estimated the split-half reliability for each of the content-dimension specific IRT variability traits and results showed high split-half reliabilities ($r^*_{tt}$ = .86 to .89; see Table 7). However, the split-half reliabilities from the same content dimension were not much higher than crossing a

[4]Note also that Pseudoitem I and Psseudoitem III only cover a subsection of the distribution so that items and persons with certain combinations of extreme latent traits only have a limited number of observations for one or both of these pseudoitems. In other IRT models, it is not uncommon to observe no association or even reversed associations among subsections of the continuum because of this phenomenon (Wetzel & Carstensen, 2014).

particular dimension with any of the other content dimension ($r^*_{tt}$ = .67 to .89; average $r^*_{tt}$ = .77). Results for Study 2 revealed that the separate IRT variability dimensions on the basis of the 10 items for each scale were again correlated but also showed unsystematic correlations with the content dimensions (see Table 6). These findings suggest that it is advisable to use more than 10 items to get stable estimates for IRT variability. Overall, our results confirm earlier findings in the literature suggesting that within-person variability largely generalizes across different content dimensions and can be measured with high reliability but requires a somewhat larger number of items than content personality traits.

**Relationship with Means and Within-Person *SD*s Across Ratings of Personality Descriptions**

To further study the characteristics of the TVTM, we examined the correlation between the latent scores from the TVTM and non-IRT operationalizations of the same constructs. As shown in Tables 6 and 7, the raw mean scores for the traits were highly correlated (average $r$ = .94 in Study 1 and .89 in Study 2) with the IRT scores for the same traits from the TVTM.

We were also interested in the degree to which IRT variability in personality ratings of personality descriptions is correlated with within-person *SD* values across items as an intuitive measure of within-person variability. In the introduction of this article, we described several challenges for studying within-person variability including the issues around the interpretation of within-person *SD*s across items with different content. However, it would nevertheless be reassuring when IRT variability would be highly correlated with this intuitive measure of within-person variability. The results in Tables 6 and 7 indeed suggest a strong relationship ($r$ = .93 in Study 1, and $r$ = .54 in Study 2) between IRT variability and the within-person *SD* values across personality items. IRT variability thus is strongly correlated with this intuitive measure of within-person variability.

**Relationship with Within-Person *SD*s in a Diary Study**

Table 7 provides correlations between the latent scores from the TVTM and *M*s and *SD*s for the personality ratings from the diary study 20 months later. As indicated by Table 7, the correlations between the latent scores for the traits and the average score across the 10 days were $r = .29$, $r = .14$, and $r = .22$ for extraversion, agreeableness, and conscientiousness, respectively. Although a considerable amount of time passed between the initial personality inventory and the diary study, these values are in a similar range as correlations between full personality inventory score and daily ratings when different items are used (Moskowitz & Zuroff, 2004) like in the present study (personality items and adjectives).

Our focus in this article was on extracting a meaningful measure of within-person variability from ratings of personality descriptions as an alternative perspective on within-person variability. The focus in existing research, in contrast, is typically on within-person variability across days. Although the two types of within-person variability are not conceptually identical, we were nevertheless interested in examining the degree to which these two perspectives on within-person variability may be related. As shown in Table 7, the correlation between IRT variability and the within-person *SD* across the daily scores 20 months later was $r = .20$. This correlation was broadly in the same range as the correlations between directional responding and the mean scores in the diary part of the study and indicates that the variability-across-time and the variability-across-descriptions perspectives on within-person variability overlap but are not identical.

We were also interested in the degree to which IRT variability across personality descriptions is capable of predicting the amount of within-person variability in adjective ratings of days and within-person variability across both days and adjective ratings in a diary format as intuitive measures of daily within-person variability and overall diary variability. As indicated by

Table 7, the correlation with the within-person *SD* across all ratings was $r$ = .38, and the average correlation with the within-person *SD* across ratings for a particular day was mean $M_r$ = .29 suggesting that IRT variability also predicts these types of within-person variability.

**Other Correlates**

In the next step, we studied the degree to which IRT variability is observable by others using the data of Study 1. Results are also provided in Table 6 and suggest that observers can rate IRT variability to some degree. The correlation between the TVTM from the self-report and from the observer-report personality inventory was $r$ = .38. However, IRT variability is not as observable as the directional traits ($r$ = .58 to $r$ = .70). One potential explanation is that judging variability may be more difficult for observers because it may require a larger behavior sample. Typically, even observers who know a target well only know the person in a limited subset of the person's environments.

Finally, we examined the degree to which IRT variability scores are distinct from person unreliability or person fit indices. The results suggested that IRT variability is not systematically correlated with person fit on the directional traits (operationalized by the recommended $l_z^*$ index, see Snijders, 2001). This finding suggests that IRT variability captures a construct that is distinct from person misfit or person unreliability. A person can thus both be high or low in IRT variability and high or low in person misfit.

## Discussion

Variability in a response pattern can stem from several different sources. As early as 1955, Fiske and Rice asked "Can we partial out from the conventional error variance of psychometrics a component of variance over time which is associated with the individual? … Are there variability factors analogous to the well-known factors of level scores in mental abilities, interests, and personality?" (p. 217). A key methodological question for research on meaningful

within-person variability has accordingly long been whether and how meaningful within-person variability can be separated from other types of variability.

In this article, we examined an IRT approach—the trait variability tree model (TVTM)—that seeks to address some of the conceptual and methodological challenges in the within-person literature. The approach allows researchers to clearly separate meaningful within-person variability from other sources of variance like measurement error and person misfit. The approach is useful because it allows researchers to extract meaningful within-person variability from widely available Likert-scale ratings of personality descriptions. Our empirical results suggest that IRT variability across ratings of personality descriptions has high split-half reliabilities, and is a variability construct that is strongly associated with the within-person *SD* across personality descriptions (an intuitive but potentially confounded measure of within-person variability). Our findings also suggest that IRT variability generalizes across content dimensions. This idea is in line with earlier results in the literature suggesting a positive manifold (Baird et al., 2006) or general oscillation factor (Spearman & Jones, 1950). In interpreting IRT variability, it is important to be aware that IRT variability is distinct from the content of the items and thus not conceptually related to the concept of higher-order personality traits above the FFM/HEXACO dimensions (e.g., Ashton, Lee, Goldberg, & de Vries, 2009). Our results also suggest that IRT variability across personality descriptions is associated with within-person variability across days. These findings are in line with the idea that the variability across personality descriptions and the variability across time perspectives on within-person variability overlap but are not identical and thus may complement each other. Furthermore, IRT variability also shows moderate correlations across observer reports and is unrelated to IRT person fit or unreliability coefficients.

**Implications**

Some readers may suggest that the described IRT approach introduces more complexity into the study of variability concepts in personality research. An IRT approach is likely almost always methodologically more complex than simply estimating means and *SD*s. However, the described tree modeling approach and tree models in general are not very complex psychometric models when one compares these models to other approaches commonly used in assessment research (R code provided in the Appendix). For instance, an ordinary 2PL IRT model that is functionally equivalent to a unidimensional confirmatory factor analysis model has a considerably higher number of parameters than the TVTM. Furthermore, while it may be true that the approach introduces additional complexity at the level of the statistical methods, the current approach also has the potential to broaden research on variability concepts in personality. Specifically, the approach has two practical implications for personality assessment on meaningful within-person variability.

First, the approach allows researchers and practitioners to more easily gather a measure of within-person variability. Researchers and test developers can apply the TVTM to ordinary personality data like we did in our Study 1 and Study 2 and routinely extract a meaningful variability trait from this type of data (IRT variability). This additional information can then be used in personality assessment without the need to collect additional data.

Second, the approach makes it possible to explicitly incorporate variability in test development and test administration because the psychometric definition of IRT variability allows researchers to assign item parameters to the variability processes at the level of the items. These item parameters can be used for item selection in both test development and adaptive testing. Especially, adaptive testing for variability processes was previously not possible because of the missing IRT model for the within-person *SD*.

In addition to the described practical implications, the suggested approach also has several theoretical implications. Most importantly, the method clarifies the conceptual relationships between several previously disjointed constructs and literatures. First, the described methodological framework integrates research on within-person variability in personality with IRT. Second, our article established a clear theoretical link between extant research on response styles like midpoint responding in personality measures using the three-process model (Böckenholt & Meiser, 2017; Zettler et al., 2016) and the personality literature on within-person variability. While early literature on response styles has typically interpreted response style as a sort of unwanted confounder, this perspective has recently started to shift and researchers now increasingly interpret response styles as substantive individual difference variables (Wetzel, Lüdtke, Zettler, & Böhnke, 2016). The conceptual link to IRT variability contributes to this perspective. Third, we suggest a way that allows researchers to separate research on IRT person misfit from within-person variability research. Until now, these two literatures existed in parallel and it was not conceptually clear to what degree research studying the correlates of careless responding or misfit (Schmitt et al., 1999) and research studying within-person *SD*s were studying the same theoretical construct. IRT and the TVTM provide a clear conceptual distinction between person misfit as a person-specific unreliability construct and IRT variability as a substantive variability construct. This distinction between variability constructs is somewhat analog to the distinction between dependability and stability for test-retest constructs (Watson, 2004). Dependability implies the absence of unwanted test-retest fluctuations because of measurement error (low dependability) and (in)stability captures test-retest fluctuations because of substantive construct change.

**Limitations and Future Directions**

Our article has several noteworthy limitations. One limitation is that the scope of the

presented empirical analyses is largely limited to the basic characteristics of the TVTM. Future research could especially explore the usefulness of IRT variability and possibly also specific IRT variability dimensions for predicting important outcome criteria in applied settings. For instance, variability concepts and dynamic change have a prominent role in the definition and description of several clinical disorders (e.g., personality pathology symptoms, see Wright & Simms, 2016). We believe that IRT variability in personality could be useful in predicting criteria of this type. Furthermore, future research could also further explore IRT variability in ratings of personality descriptions and within-person variability across time by combining the two different types of within-person variability in joint designs and analyses. As a starting point and on the basis of the suggestion of a reviewer, we applied an extended multilevel version of the TVTM with measurement occasions nested in person to the diary data in Study 2. Our goal was to study whether IRT variability in the daily personality adjective ratings was as stable across days as the content traits. This research question is of interest because constructs should be dependable in the sense that change over very short periods should not be unreasonably high (Watson, 2004). The multilevel TVTM model adds an additional random effect $\xi$ capturing daily observations for a particular person for each construct in addition to the existing $\theta$ effects for the person. Results estimated with intraclass correlations (variance between persons / [variance between persons + variance within persons]) revealed that 71, 35, 95, and 52 percent of the variance in IRT variability, extraversion, agreeableness, and conscientiousness, respectively, was stable across days. The IRT variability trait estimated from the diary study also showed stability in the sense that it was substantially correlated with the IRT variability trait estimated from the personality descriptions 20 month earlier ($r$ = .40).

A second limitation of this article is that the described approach is limited to Likert-type response data. Future research could study the measurement of IRT variability concepts with

other types of rating scales. IRT tree models are relatively flexible and it should thus be possible to adapt existing models for other types of rating scales to measure variability traits (Böckenholt, 2012; Böckenholt & Meiser, 2017; Lievens et al., 2017). Another potential goal for future research could be to explore ways to incorporate variability concepts into other types of response formats used in personality psychology like, for instance, forced-choice measures.

## References

Allport, G. W. (1937). *Personality: A psychological interpretation.* New York: Holt.

Allport, G. W. (1965). *Letters from Jenny*. San Diego, CA: Harcourt.

Ashton, M. C., Lee, K., Goldberg, L. R., & de Vries, R. E. (2009). Higher Order Factors of Personality: Do They Exist? *Personality and Social Psychology Review*, *13*, 79–91. https://doi.org/10.1177/1088868309338467

Baird, B. M., Le, K., & Lucas, R. E. (2006). On the nature of intraindividual personality variability: Reliability, validity, and associations with well-being. *Journal of Personality and Social Psychology*, *90*, 512–527. https://doi.org/10.1037/0022-3514.90.3.512

Bates, D. M., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models using lme4. *Journal of Statistical Software*. Retrieved from http://arxiv.org/abs/1406.5823

Bem, D. J., & Allen, A. (1974). On predicting some of the people some of the time: The search for cross-situational consistencies in behavior. *Psychological Review*, *81*, 506–520. https://doi.org/10.1037/h0037130

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods*, *17*, 665–678. https://doi.org/10.1037/a0028111

Böckenholt, U., & Meiser, T. (2017). Response style analysis with threshold and multi-process IRT models: A review and tutorial. *British Journal of Mathematical and Statistical*

*Psychology*, *70*, 159–181. https://doi.org/10.1111/bmsp.12086

Chaplin, W. L., & Goldberg, L. R. (1985). A failure to replicate the Bem and Allen study of individual differences in cross-situational consistency. *Journal of Personality and Social Psychology*, *47*, 1074–1090.

Conner, T. S., Tennen, H., Fleeson, W., & Barrett, L. F. (2009). Experience sampling methods: A modern idiographic approach to personality research. *Social and Personality Psychology Compass*, *3*(3), 292–313. https://doi.org/10.1111/j.1751-9004.2009.00170.x

Costa, P. T., & McCrae, R. R. (1992). *NEO PI-R professional manual.* Odessa, FL: Psychological Assessment Resources, Inc.

Costa, P. T., & McCrae, R. R. (1995). Domains and facets: hierarchical personality assessment using the revised NEO personality inventory. *Journal of Personality Assessment*, *64*(1), 21–50. https://doi.org/10.1207/s15327752jpa6401_2

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*, 533–559. https://doi.org/10.1007/s11336-008-9092-x

De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, *48*(1), 1–18. https://doi.org/10.18637/jss.v048.c01

Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007). Estimating the Multilevel Rasch Model: With the lme4 package. *Journal of Statistical Software*, *20*, 1–18. https://doi.org/10.1111/j.1467-9868.2007.00600.x

Eid, M., & Diener, E. (1999). Intraindividual variability in affect: Reliability, validity, and personality correlates. *Journal of Personality and Social Psychology*, *76*, 662–676. https://doi.org/10.1037//0022-3514.76.4.662

Eysenck, H. J. (1953). *The structure of human personality*. London: Methuen.

Eysenck, H. J., & Eysenck, M. (1985). *Personality and individual differences: A natural science*

*approach*. New York: Plenum Press.

Fiske, D. W., & Rice, L. (1955). Intra-individual response variability. *Psychological Bulletin*, *52*, 217–250. https://doi.org/10.1037/h0045276

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*, 1011–1027. https://doi.org/10.1037//0022-3514.80.6.1011

Fleeson, W., & Jayawickreme, E. (2015). Whole trait theory. *Journal of Research in Personality*, *56*, 82–92. https://doi.org/10.1016/j.jrp.2014.10.009

Fleisher, M. S., Woehr, D. J., Edwards, B. D., & Cullen, K. L. (2011). Assessing within-person personality variability via frequency estimation: More evidence for a new measurement approach. *Journal of Research in Personality*, *45*, 535–548. https://doi.org/10.1016/j.jrp.2011.06.009

Flügel, J. C. (1929). Practice, fatigue, and oscillation. *British Journal of Psychology*, *7*, Monograph Supplement, 13.

Goldberg, L. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality Psychology in Europe, Vol. 7* (pp. 7–28). Tilburg, The Netherlands: Tilburg University Press.

Greene, W. H. (2012). *Econometric analysis* (7th ed.). Essex, UK: Pearson.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hollingworth, H. L. (1925). Correlations of achievement within an individual. *Journal of Experimental Psychology*, *8*, 190–208. https://doi.org/10.1037/h0065414

Jeon, M., & De Boeck, P. (2016). A generalized item response tree model for psychological

assessments. *Behavior Research Methods*, *48*, 1070–1085. https://doi.org/10.3758/s13428-015-0631-y

Kehr, T. (1916). Versuchsanordnung zur experimentellen Untersuchung einer kontinuierlichen Aufmerksamkeitsleistung. *Zeitschrift Für Angewandte Psychologie*, *11*, 465–479.

Lang, J. W. B., Tackett, J. L., & Zettler, I. (2017). Open peer commentary: Utilizing advanced psychometric methods in research on trait expression across situations. *European Journal of Personality*, *31*, 464–465. https://doi.org/10.1002/per

Lee, K., & Ashton, M. C. (2006). Further assessment of the HEXACO Personality Inventory: Two new facet scales and an observer report form. *Psychological Assessment*, *18*, 182–191. https://doi.org/10.1037/1040-3590.18.2.182

Lee, K., & Ashton, M. C. (2008). The HEXACO personality factors in the indigenous personality lexicons of English and 11 other languages. *Journal of Personality*, *76*(5), 1001–1054. https://doi.org/10.1111/j.1467-6494.2008.00512.x

Lievens, F. (2017). Author's response: Integrating situational judgment tests and assessment centre exercises into personality research: Challenges and further opportunities. *European Journal of Personality*, *31*, 487–502. https://doi.org/10.1002/per

Lievens, F., Lang, J. W. B., De Fruyt, F., Corstjens, J., Van De Vijver, M., & Bledow, R. (2018). The predictive power of people's intra-individual variability across situations: Implementing whole trait theory in assessment. *Journal of Applied Psychology*.

Magis, D., Raiche, G., & Beland, S. (2012). A didactic presentation of Snijders's $l_z^*$ index of person fit with emphasis on response model selection and ability estimation. *Journal of Educational and Behavioral Statistics*, *37*, 57–81. https://doi.org/10.3102/1076998610396894

McFadden, D. (2001). Economic choices. *American Economic Review*, *91*, 351–378.

https://doi.org/10.1257/aer.91.3.351

Meijer, R. R., & Sijtsma, K. (2001). Methodology Review: Evaluating Person Fit. *Applied Psychological Measurement*, *25*, 107–135. https://doi.org/10.1177/01466210122031957

Meijer, R. R., & Tendeiro, J. N. (2012). The use of the lz and lz* person-fit statistics and problems derived from model misspecification. *Journal of Educational and Behavioral Statistics*, *37*, 758–766. https://doi.org/10.3102/1076998612466144

Mischel, W. (1968). *Personality and Assessment*. New York, NY: Wiley.

Moskowitz, D. S., & Zuroff, D. C. (2004). Flux, pulse, and spin: Dynamic additions to the personality lexicon. *Journal of Personality and Social Psychology*, *86*, 880–893. https://doi.org/10.1037/0022-3514.86.6.880

Paulhus, D. L., & Martin, C. L. (1988). Functional flexibility: A new conception of interpersonal flexibility. *Journal of Personality and Social Psychology*, *55*(1), 88–101. https://doi.org/10.1037/0022-3514.55.1.88

Pervin, L. A. (1994). Further reflections on current trait theory. *Psychological Inquiry*, *5*, 169–178.

Pulakos, E. D., Arad, S., Donovan, M. A., & Plamondon, K. E. (2000). Adaptability in the workplace: Development of a taxonomy of adaptive performance. *Journal of Applied Psychology*, *85*, 612–624. https://doi.org/10.1037/0021-9010.85.4.612

R Core Team. (2014). R: A Language and Environment for Statistical Computing [Version 3.1.1]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org/

Reise, S. P. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement*, *19*, 213–229. https://doi.org/10.1177/014662169501900301

Schmitt, N., Chan, D., Sacco, J. M., McFarland, L. A., & Jennings, D. (1999). Correlates of person fit and effect of person fit on test validity. *Applied Psychological Measurement*, *23*, 41–53. https://doi.org/10.1177/01466219922031176

Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika*, *66*, 331–342. https://doi.org/10.1007/BF02294437

Soto, C. J., & John, O. P. (2009). Ten facet scales for the Big Five Inventory: Convergence with NEO PI-R facets, self-peer agreement, and discriminant validity. *Journal of Research in Personality*, *43*(1), 84–90. https://doi.org/10.1016/j.jrp.2008.10.002

Spearman, C., & Jones, L. W. (1950). *Human ability*. London, UK: Macmillan.

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology*, *91*, 25–39. https://doi.org/10.1037/0021-9010.91.1.25

Walton, R. D. (1936). Relations between amplitude of oscillations in short period efficiency and steadiness of character. *British Journal of Psychology, XXVII*.

Watson, D. (2004). Stability versus change, dependability versus error: Issues in the assessment of personality over time. *Journal of Research in Personality*, *38*(4), 319–350. https://doi.org/10.1016/j.jrp.2004.03.001

Werner, P. D., & Pervin, L. A. (1986). The content of personality inventory items. *Journal of Personality and Social Psychology*, *51*(3), 622–8. https://doi.org/10.1037/0022-3514.51.3.622

Wetzel, E., & Carstensen, C. H. (2014). Reversed Thresholds in Partial Credit Models: A Reason for Collapsing Categories? *Assessment*, *21*(6), 765–774. https://doi.org/10.1177/1073191114530775

Wetzel, E., Lüdtke, O., Zettler, I., & Böhnke, J. R. (2016). The Stability of Extreme Response Style and Acquiescence Over 8 Years. *Assessment*, *23*(3), 279–291. https://doi.org/10.1177/1073191115583714

Winter, D. G., John, O. P., Stewart, A. J., Klohnen, E. C., & Duncan, L. E. (1998). Traits and motives: toward an integration of two traditions in personality research. *Psychological Review*, *105*(2), 230–250. https://doi.org/10.1037//0033-295x.105.2.230

Wright, A. G. C., & Simms, L. J. (2016). Stability and fluctuation of personality disorder features in daily life. *Journal of Abnormal Psychology*, *125*, 641–656. https://doi.org/10.1037/abn0000169

Zettler, I., Lang, J. W. B., 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*, 461–472. https://doi.org/10.1111/jopy.12172

## Appendix: R Code For Fitting the Trait Variability Tree Model

The tree variability tree model (TVTM) builds on Böckenholt's (2012) three-process IRT tree model for Likert-scale data which in turn is based on earlier work in economics (e.g., see Greene, 2012; McFadden, 2001).

```
# The data.frame with the items for each dimension
# after each other and recoded
# the assumption is that all dimensions have the same number
# of items

mydataset<-bigfive
dimensions<-c("A","C","E","N","O")

library(irtrees)
library(lme4)

# the model specification
mapping <- cbind(c(1,1,0,1,1),
             c(0,0,NA,1,1),
             c(0,1,NA,1,0))

#prepare the data
edat<-dendrify(as.matrix(mydataset),mapping)

#further define dimensions and pseudoitems
edat$node1<-ifelse(edat$node=="node1",1,0)
edat$node2<-ifelse(edat$node=="node2",1,0)
edat$node3<-ifelse(edat$node=="node3",1,0)
edat$itemn<-as.numeric(sub("i","",as.character(edat$item)))
edat$dim<-cut(edat$itemn,breaks=length(dimensions),dimensions)

model1 <- glmer(value ~ 0+ node + (0+node | item) +
       (0+node2:dim+I(node1+node3*(-1))| person),
       family = binomial("probit"), data = edat,
       control=glmerControl(optimizer="nloptwrap",calc.deriv=F))
summary(model1)
```

Table 1
Responses by a Person to Four Items at Four Different Points in Time

| Item | Time 1 | Time 2 | Time 3 | Time 4 | *SD* across time |
|---|---|---|---|---|---|
| Approached unfamiliar people | 5 | 4 | 5 | 5 | |
| Look the lead on something in a group context | 1 | 1 | 2 | 1 | |
| Acted socially | 5 | 4 | 5 | 5 | |
| Glanced at something horrific | 1 | 2 | 1 | 1 | |
| Scale | 12 | 11 | 13 | 12 | .82 |

Table 2
Responses by Three Persons to Four Extraversion Items

| Item | Person 1 | Person 2 | Person 3 |
|---|---|---|---|
| 1. I like to go to parties | 2 | 5 | 1 |
| 2. I like horror movies | 3 | 5 | 3 |
| 3. I am a social person | 2 | 5 | 2 |
| 4. In meetings, I usually take the lead | 3 | 5 | 4 |
| *M* | 2.5 | 5 | 2.5 |
| *SD* | 0.6 | 0 | 1.3 |

Table 3

Category Probabilities for the Trait Variability Tree Model

| Response option | Pseudoitem I | Pseudoitem II | Pseudoitem III |
|---|---|---|---|
| Coding | | | |
| strongly disagree | 1 | 0 | 0 |
| disagree | 1 | 0 | 1 |
| neutral | 0 | – | – |
| agree | 1 | 1 | 1 |
| strongly agree | 1 | 1 | 0 |
| Probabilities | | | |
| Pr($y_{jk}$ = strongly disagree) = | $\Phi\left(\gamma_j^{(I)} + \theta_k^{(V)}\right)$ | $\left[1 - \Phi\left(\gamma_j^{(II)} + \theta_k^{(IIA, IIB, \ldots IIX)}\right)\right]$ | $[1 - \Phi\left(\gamma_j^{(III)} - \theta_k^{(V)}\right)$ |
| Pr($y_{jk}$ = disagree) = | $\Phi\left(\gamma_j^{(I)} + \theta_k^{(V)}\right)$ | $\left[1 - \Phi\left(\gamma_j^{(II)} + \theta_k^{(IIA, IIB, \ldots IIX)}\right)\right]$ | $\Phi\left(\gamma_j^{(III)} - \theta_k^{(V)}\right)$ |
| Pr($y_{jk}$ = neutral) = | $1 - \Phi(\gamma_j^{(I)} + \theta_k^{(I)})$ | | |
| Pr($y_{jk}$ = agree) = | $\Phi\left(\gamma_j^{(I)} + \theta_k^{(V)}\right)$ | $\Phi\left(\gamma_j^{(II)} + \theta_k^{(IIA, IIB, \ldots IIX)}\right)$ | $\Phi\left(\gamma_j^{(III)} - \theta_k^{(V)}\right)$ |
| Pr($y_{jk}$ = strongly agree) = | $\Phi\left(\gamma_j^{(I)} + \theta_k^{(V)}\right)$ | $\Phi\left(\gamma_j^{(II)} + \theta_k^{(IIA, IIB, \ldots IIX)}\right)$ | $[1 - \Phi\left(\gamma_j^{(III)} - \theta_k^{(V)}\right)$ |

Table 4

Study 1: Model Estimates for the Trait Variability Tree Model

| | | Latent correlations | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Estimate | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| Model 1 | | | | | | | | | | | | |
| Indifference | 0.85 | | | | | | | | | | | |
| Direction | 0.46 | | | | | | | | | | | |
| Intensity | 0.72 | | | | | | | | | | | |
| Item variance | | | | | | | | | | | | |
| 1. $\sigma_{\text{pseudoitem I}}$ | 0.23 | | | | | | | | | | | |
| 2. $\sigma_{\text{pseudoitem II}}$ | 0.71 | .26 | | | | | | | | | | |
| 3. $\sigma_{\text{pseudoitem III}}$ | 0.30 | -.66 | -.33 | | | | | | | | | |
| Person variance | | | | | | | | | | | | |
| 1. $\sigma_{\text{IRT variability (VAR)}}$ | 0.32 | | | | | | | | | | | |
| 2. $\sigma_{\text{honesty-humility (HH)}}$ | 0.77 | -.17 | | | | | | | | | | |
| 3. $\sigma_{\text{emotionality (EM)}}$ | 0.74 | -.10 | .26 | | | | | | | | | |
| 4. $\sigma_{\text{eXtraversion (EX)}}$ | 0.74 | -.07 | -.13 | -.16 | | | | | | | | |
| 5. $\sigma_{\text{agreeableness (A)}}$ | 0.70 | -.02 | .38 | .07 | .16 | | | | | | | |
| 6. $\sigma_{\text{conscientiousness (C)}}$ | 0.68 | -.11 | .29 | -.03 | .18 | .13 | | | | | | |
| 7. $\sigma_{\text{openness (O)}}$ | 0.70 | .06 | .01 | .00 | .20 | .07 | .01 | | | | | |
| -2LogLikelihood | 289,844 | | | | | | | | | | | |
| Model 2 | | | | | | | | | | | | |
| Indifference | 0.87 | | | | | | | | | | | |
| Direction | 0.46 | | | | | | | | | | | |
| Intensity | 0.76 | | | | | | | | | | | |
| Item variance | | | | | | | | | | | | |
| 1. $\sigma_{\text{pseudoitem I}}$ | 0.24 | | | | | | | | | | | |
| 2. $\sigma_{\text{pseudoitem II}}$ | 0.71 | .28 | | | | | | | | | | |
| 3. $\sigma_{\text{pseudoitem III}}$ | 0.31 | -.66 | -.31 | | | | | | | | | |
| Person variance | | | | | | | | | | | | |
| 1. $\sigma_{\text{IRT variability, HH}}$ | 0.43 | | | | | | | | | | | |
| 2. $\sigma_{\text{IRT variability, EM}}$ | 0.41 | .60 | | | | | | | | | | |
| 3. $\sigma_{\text{IRT variability, EX}}$ | 0.45 | .49 | .66 | | | | | | | | | |
| 4. $\sigma_{\text{IRT variability, A}}$ | 0.41 | .53 | .72 | .62 | | | | | | | | |
| 5. $\sigma_{\text{IRT variability, C}}$ | 0.40 | .50 | .65 | .60 | .78 | | | | | | | |
| 6. $\sigma_{\text{IRT variability, O}}$ | 0.36 | .53 | .62 | .52 | .69 | .65 | | | | | | |
| 7. $\sigma_{\text{HH}}$ | 0.78 | .31 | -.17 | -.24 | -.32 | -.21 | -.24 | | | | | |
| 8. $\sigma_{\text{EM}}$ | 0.74 | -.02 | .09 | -.11 | -.12 | -.13 | -.20 | .26 | | | | |
| 9. $\sigma_{\text{EX}}$ | 0.76 | -.20 | -.17 | .41 | -.15 | -.09 | -.11 | -.14 | -.17 | | | |
| 10. $\sigma_{\text{A}}$ | 0.70 | .19 | -.05 | .12 | -.25 | -.13 | -.04 | .37 | -.07 | .16 | | |
| 11. $\sigma_{\text{C}}$ | 0.68 | .02 | -.27 | -.09 | -.11 | .08 | -.17 | .28 | -.03 | .17 | .13 | |
| 12. $\sigma_{\text{O}}$ | 0.70 | -.01 | .03 | .11 | .05 | .03 | -.12 | .01 | -.00 | .20 | .07 | .01 |
| -2LogLikelihood | 287,000 | | | | | | | | | | | |

$k$ = 577 respondents; $j$ = 192 items; $n$ = 284,714 observations;

Table 5

Study 2: Model Estimates for the Trait Variability Tree Model

| | | Latent correlations | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Estimate | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Model 1 | | | | | | | | | | |
| Indifference | 1.39 | | | | | | | | | |
| Direction | 1.16 | | | | | | | | | |
| Intensity | 1.05 | | | | | | | | | |
| Item variance | | | | | | | | | | |
| 1. $\sigma_{\text{pseudoitem I}}$ | 0.34 | | | | | | | | | |
| 2. $\sigma_{\text{pseudoitem II}}$ | 0.88 | .60 | | | | | | | | |
| 3. $\sigma_{\text{pseudoitem III}}$ | 0.31 | -.80 | -.30 | | | | | | | |
| Person variance | | | | | | | | | | |
| 1. $\sigma_{\text{IRT variability}}$ | 0.45 | | | | | | | | | |
| 2. $\sigma_{\text{extraversion (EX)}}$ | 1.27 | .19 | | | | | | | | |
| 3. $\sigma_{\text{agreeableness (A)}}$ | 0.88 | .05 | .39 | | | | | | | |
| 4. $\sigma_{\text{conscientiousness (C)}}$ | 1.14 | .23 | .05 | .27 | | | | | | |
| 5. $\sigma_{\text{neuroticism (N)}}$ | 1.49 | -.12 | .25 | .25 | -.07 | | | | | |
| 6. $\sigma_{\text{openness (O)}}$ | 0.96 | .06 | .16 | .06 | -.19 | .32 | | | | |
| -2LogLikelihood | 23,270 | | | | | | | | | |
| Model 2 | | | | | | | | | | |
| Indifference | 1.39 | | | | | | | | | |
| Direction | 1.16 | | | | | | | | | |
| Intensity | 1.05 | | | | | | | | | |
| Item variance | | | | | | | | | | |
| 1. $\sigma_{\text{pseudoitem I}}$ | 0.36 | | | | | | | | | |
| 2. $\sigma_{\text{pseudoitem II}}$ | 0.88 | .62 | | | | | | | | |
| 3. $\sigma_{\text{pseudoitem III}}$ | 0.33 | -.79 | -.23 | | | | | | | |
| Person variance | | | | | | | | | | |
| 1. $\sigma_{\text{IRT variability, EX}}$ | 0.78 | | | | | | | | | |
| 2. $\sigma_{\text{IRT variability, A}}$ | 0.85 | .50 | | | | | | | | |
| 3. $\sigma_{\text{IRT variability, C}}$ | 0.78 | .38 | .50 | | | | | | | |
| 4. $\sigma_{\text{IRT variability, N}}$ | 0.62 | .25 | .34 | .32 | | | | | | |
| 5. $\sigma_{\text{IRT variability, O}}$ | 0.73 | .44 | .24 | .33 | .32 | | | | | |
| 6. $\sigma_{\text{EX}}$ | 1.37 | .59 | .19 | .00 | .05 | -.03 | | | | |
| 7. $\sigma_{\text{A}}$ | 1.02 | .06 | .60 | .08 | .10 | -.43 | .35 | | | |
| 8. $\sigma_{\text{C}}$ | 1.20 | .00 | .35 | .51 | -.06 | -.12 | .06 | .27 | | |
| 9. $\sigma_{\text{N}}$ | 1.48 | -.04 | -.17 | -.23 | .15 | -.08 | .24 | .21 | -.07 | |
| 10. $\sigma_{\text{O}}$ | 0.98 | .20 | -.08 | -.22 | .00 | .47 | .17 | -.05 | -.20 | .31 |
| -2LogLikelihood | 23,239 | | | | | | | | | |

*k* = 235 respondents; *j* = 50 items; *n* = 32,876 observations.

Table 6

Study 1: Characteristics and Correlates of the Latent Trait Scores

| | | | | | | | | Dimension-specific IRT variability | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | IRT variability | HH | EM | EX | A | C | O | HH | EM | EX | A | C | O |
| Split-half $r^*_{tt}$ ($r_{tt}$) | .92 (.85) | .87 (.77) | .87 (.78) | .84 (.72) | .84 (.73) | .81 (.68) | .85 (.63) | .87 (.77) | .89 (.81) | .89 (.80) | .89 (.80) | .88 (.78) | .86 (.76) |
| $r_{\text{latent trait score, raw mean score}}$ | | .92 | .96 | .92 | .96 | .94 | .96 | | | | | | |
| $r_{\text{latent trait score, } SD \text{ across all items}}$ | .93 | | | | | | | .66 | .86 | .72 | .89 | .83 | .84 |
| $r_{\text{IRT variability, person fit (Iz*),}}$ | | -.12 | -.15 | -.20 | -.18 | -.25 | -.14 | -.04 | -.11 | .03 | -.20 | -.20 | -.20 |
| Self-observer *r* | .38 | .67 | .70 | .66 | .58 | .61 | .69 | .41 | .37 | .55 | .36 | .37 | .35 |

*Note.* $k = 577$ respondents. Split-half reliabilities are for latent trait scores from the first and second half of the test. Values were estimated using the standard formula for split-half reliabilites that corrects for the split using the correction formula $r^*_{tt} = (2r_{tt})/(1+ r_{tt})$. The uncorrected correlations between the halves are provided in parentheses. All correlations are significant $p < .01$. Latent trait scores are from Model 1 in Table 4 with the exception of the domain-specific variability scores which are from Model 2 in the same table. HH = honest-humility; EM = emotionality; EX = extraversion; A = agreeableness; C = conscientiousness, O = openness.

Table 7

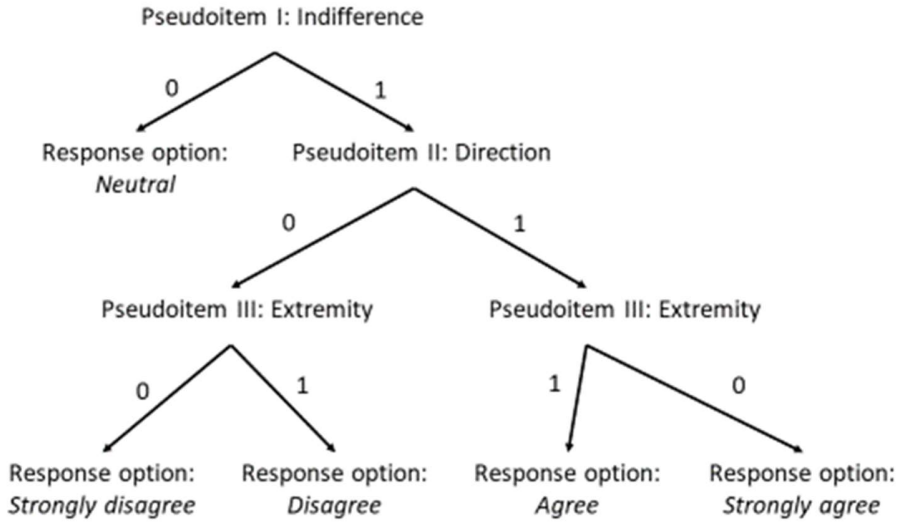
Study 2: Characteristics and Correlates of the Latent Trait Scores

| | IRT variability | EX | A | C | N | O |
|---|---|---|---|---|---|---|
| Split-half $r^*_{tt}$ ($r$) | .83 (.71**) | .79 (.65**) | .77 (.63**) | .75 (.60**) | .81 (.68**) | .74 (.59**) |
| $r_{\text{latent trait score, raw mean score}}$ | | .94** | .68** | .96** | .96** | .90** |
| $r_{\text{latent trait score, } SD \text{ across all items}}$ | .55** | | | | | |
| $r_{\text{person fit (Iz*), IRT variability}}$ | | .06 | .02 | .21** | .07 | .16 |
| Daily personality 20 months later (120 respondents) | | | | | | |
| *M* daily extraversion | .15 | .29* | .12 | .11 | -.04 | -.02 |
| *M* daily agreeableness | .10 | -.02 | .14 | .24* | .13 | -.02 |
| *M* daily conscientiousness | .09 | .00 | .13 | .22* | .03 | -.02 |
| *SD* across scale means | .20* | -.20* | -.14 | -.08 | -.15 | -.10 |
| *SD* across ratings for one day (average correlation) | .29 | .00 | .06 | .14 | -.03 | -.09 |
| *SD* across all ratings | .38** | .00 | .08 | .18* | -.05 | -.12 |

*Note.* $k$ = 235 respondents for the personality inventory and $k$ = 120 respondents for the diary study. Split-half reliabilities are for latent trait scores from the first and second half of the test. Values were estimated using the standard formula for split-half reliabilities that corrects for the split using the correction formula $r^*_{tt} = (2r_{tt})/(1+ r_{tt})$. The uncorrected correlations between the halves are provided in parentheses. EX = extraversion; A = agreeableness; C = conscientiousness; N = neuroticism; O = openness.

† $p < .10$ * $p < .05$ ** $p < .01$

A




B

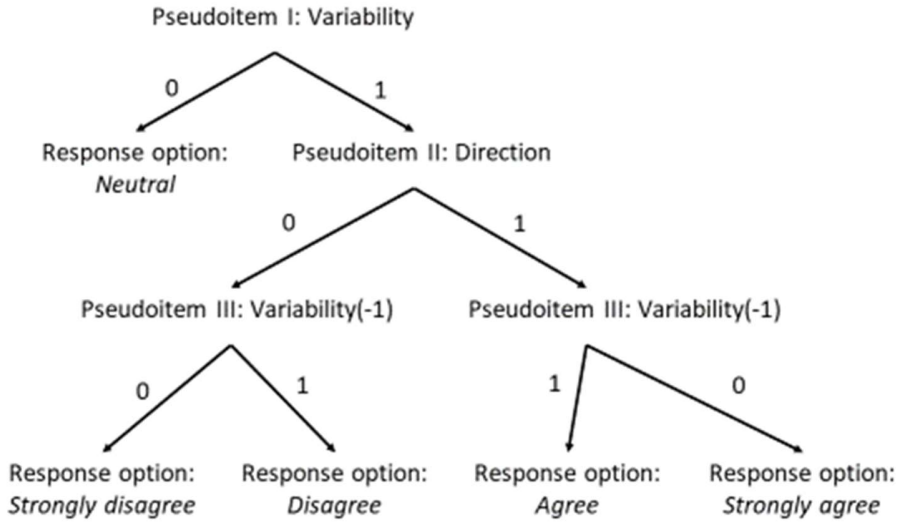



Figure 1. Three-process tree model (A) and trait variability tree model (B)