Received: 25 September 2017 | Revised: 6 March 2018 | Accepted: 21 April 2018

DOI: 10.1111/jopy.12395

ORIGINAL ARTICLE

WILEY

Are reflective models appropriate for very short scales? Proofs of concept of formative models using the Ten-Item Personality Inventory

Nils Myszkowski[1] | **Martin Storme**[2] | **Jean-Louis Tavani**[3]

[1] Pace University

[2] Université Paris Descartes

[3] Université Paris Vincennes

Correspondence
Nils Myszkowski, Department of Psychology, Pace University, 41 Park Row, New York, NY 10038.
Email: nmyszkowski@pace.edu

Abstract

Objective: Because of their length and objective of broad content coverage, very short scales can show limited internal consistency and structural validity. We argue that it is because their objectives may be better aligned with formative investigations than with reflective measurement methods that capitalize on content overlap. As proofs of concept of formative investigations of short scales, we investigate the Ten-Item Personality Inventory (TIPI).

Method: In Study 1, we administered the TIPI and the Big Five Inventory (BFI) to 938 adults and fitted a formative Multiple Indicators Multiple Causes model, which consisted of the TIPI items forming five latent variables, which in turn predicted the five BFI scores. These results were replicated in Study 2 on a sample of 759 adults, but this time with the Revised NEO Personality Inventory (NEO-PI-R) as the external criterion.

Results: The models fit the data adequately, and moderate to strong significant effects ($.37 < |\beta| < .69$, all $ps < .001$) of all five latent formative variables on their corresponding BFI and NEO-PI-R scores were observed.

Conclusions: This study presents a formative approach that we propose to be more consistent with the aims of scales with broad content and short length like the TIPI.

KEYWORDS

Five-Factor Model, formative indicators, formative MIMIC, short scales

1 | INTRODUCTION

While most scales are designed to maximize internal consistency, most very short scales, such as the Ten-Item Personality Inventory (TIPI; Gosling, Rentfrow, & Swann, 2003), emphasize "content validity considerations, resulting in low inter-item correlations" (Gosling et al., 2003, p. 516). As a result, investigating such scales with traditional psychometric tools (Cronbach's α, factor analysis), which assume (at least) congenericity, results in problematic estimates and negatively biased conclusions about the usefulness of very short scales (Ziegler, Kemper, & Kruyen, 2014). As very short scales strive "for breadth of coverage" (Gosling et al., 2003, p. 508), we here propose that they could be investigated using formative models. As opposed to reflective models, where items are considered *effects* of common latent traits—thus necessarily correlated—in formative models, items are considered *samples* of a type of behavior, and thus possibly but not necessarily correlated (Markus & Borsboom, 2013). With the TIPI's being one of the most notorious very short scales, we here propose two proof-of-concept formative investigations of the French TIPI: one that uses the Big Five Inventory (BFI; John & Srivastava, 1999) and one that uses the Revised NEO Personality Inventory (NEO-PI-R; Costa & McCrae, 1992b).

1.1 | The Ten-Item Personality Inventory

The Five-Factor Model of personality is one of the most heavily used frameworks (John & Srivastava, 1999) when it comes to studying relations between personality traits and other variables in a variety of domains of psychology. Therefore, the availability of practical and psychometrically robust measures of the Big Five personality traits is an important matter. Addressing this need, the TIPI (Gosling et al., 2003), an extremely brief measure of the Big Five personality traits, has recently gained a considerable amount of attention (its original article has been cited more than 3,000 times since 2003). It has demonstrated good convergent validity, test–retest reliability, and convergence between self- and observer ratings (Gosling et al., 2003). Its multiple translations—in Dutch (Hofmans, Kuppens, & Allik, 2008), French (Storme, Tavani, & Myszkowski, 2016), Italian (Chiorri, Bracco, Piccinno, Modafferi, & Battini, 2014), German (Muck, Hell, & Gosling, 2007), and Spanish (Renau, Oberst, Gosling, Rusiñol, & Chamarro, 2013)—have also been the object of encouraging psychometrical investigations.

Despite its satisfactory concurrent validity and test–retest reliability (Gosling et al., 2003; Storme et al., 2016), some versions of the TIPI have shown problematic internal consistency estimates (Gosling et al., 2003; Muck et al., 2007; Storme et al., 2016), as well as challenging confirmatory factor analysis (CFA) results (Muck et al., 2007; Storme et al., 2016). Researchers have notably pointed to the brevity of the measure as the main explanation (e.g., Gosling et al., 2003; Oshio, Abe, Cutrone, & Gosling, 2014; Storme et al., 2016). Further investigations have indicated that failing to account for shared method variance may impact such fit indices (Oshio et al., 2014). These investigations, however, all assume that the items of the same scale are caused by the same attribute. In the next section, we explain that it represents one approach—albeit unarguably mainstream (Bollen & Diamantopoulos, 2017)—to measurement theory, and that it can be questioned whether this approach corresponds to the construction process and objectives of the TIPI.

1.2 | Reflexive and formative indicators

Like the TIPI in its previous investigations, most measures are investigated by conceptualizing the items as caused by latent variables. These investigations, often labeled as reflective (Bollen & Diamantopoulos, 2017), are an application of the Causal Theory of Measurement (CTM), in which a construct—for example, Agreeableness—instigates the set of item responses (Markus & Borsboom, 2013). This causal structure is implied in the most commonly used psychometric analyses, notably classical test theory–related analyses (e.g., Cronbach's α), item response theory, and factor analysis (Bollen & Diamantopoulos, 2017). In the CTM framework, individuals respond differently to the items because of their common latent attribute, and the relation between item responses and the construct is of explanatory nature. Estimating the construct thus consists of estimating the cause from its effects. Applied to our example, with the TIPI items, CTM states that one's latent Agreeableness *causes* both describing oneself as "Sympathetic, warm" and as (not) "Critical, quarrelsome." This theory thus implies that because the two items have a common cause, their responses should be correlated.

However, the reflective framework contrasts with the formative conceptualization of constructs. Formative models are applications of Behavior Domain Theory (BDT), in which "constructs are conceptualized in terms of domains of behavior, and item responses are considered samples from this domain" (Markus & Borsboom, 2013, p. 54). Under this framework, instead of being of a causal nature, the relation between the item responses and the construct is a sample–population relation. Estimating the construct under this theory consists of an inference based on a generalization in the population from a sample. Applied to our example, BDT states that because one describes oneself as "Sympathetic, warm" and as (not) "Critical, quarrelsome," we can *generalize* and infer that one behaves with Agreeableness in general. This theory assumes that the two items form a representative sample of agreeable behavior—but not that their responses are necessarily related. Thus, in this framework, for a given latent formative variable, correlations between the items may or may not exist, and investigations that rely on inter-item correlations (e.g., internal consistency estimates) become irrelevant. Instead, researchers focus on whether a representative sample of behavior is available from the items (Markus & Borsboom, 2013), and on the extent to which the construct formed mediates the relation between the item responses and external criteria, using characteristically the Multiple Indicators Multiple Causes (MIMIC) model (Diamantopoulos & Winklhofer, 2001; Joreskog & Goldberger, 1975).

1.3 | Are the Big Five reflective by nature?

The Five-Factor Model of personality was especially discovered using reflective measurement techniques, particularly exploratory factor analysis (EFA; Cattell, 1943; Goldberg, 1990; McCrae & John, 1992). Indeed, the Big Five—Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness—were identified as general traits that underlie "smaller" elements. Because of this "historical" reflectivity, and of typical go-to psychometrical procedures (e.g., Cronbach's α, CFA, EFA), the Big Five personality traits are traditionally assumed to cause their facets and indicators, regardless of the measure.

Nevertheless, recent measurement theory advances (Markus & Borsboom, 2013) explain that the approach to measurement—CTM or BDT—may ultimately be decided by the test developer when building the measure, not by which construct is measured. If the objective underlying test development is to use effects of Agreeableness for its estimation, then an appropriate set of items consists of items whose responses are effectively all *caused* by a common attribute of Agreeableness. If, instead, the objective behind writing items is to represent Agreeableness to generalize to the entire domain of agreeable behavior, then a good set of items consists of items responses that *sample*—or "cover"—agreeable behavior as broadly as possible. Therefore, the construct studied itself may not necessitate or call for one approach or the other.

Consequently, even though the theory upon which the TIPI was built implied reflective procedures, this should not automatically indicate that the relation between the constructs of the TIPI and its items is automatically of a reflective nature. Applying the previously suggested (Bainter & Bollen, 2014; Bollen & Diamantopoulos, 2017) mental experiment to identify formative constructs, and using Neuroticism as an example construct, we could imagine (a) that the individual characteristics of frequently getting very angry and of feeling self-conscious are, for example, two characteristics that both correspond to the conceptual unity of Neuroticism (in that they correspond to the definition of the construct); (b) that an increase in one of these characteristics would imply an increase in Neuroticism; and (c) that these two characteristics, although not necessarily independent, are not automatically expected to be found in the same individuals. Therefore, it appears from this mental experiment that it may be possible for a Big Five trait to be estimated with a formative framework.

1.4 | Is the TIPI formatively conceptualized?

As was pointed out by previous research (Bollen & Diamantopoulos, 2017; Diamantopoulos & Winklhofer, 2001), failure to recognize a formative or a reflexive model leads to inappropriate use of statistical procedures (typically, Cronbach's α and factor analysis) and consequently leads to discarding valid measures based on such inappropriate procedures. The question of investigating the TIPI and its quality with appropriate techniques is thus an important one.

Probably due to the fact that it is a fairly recent debate in psychological measurement theory (Bollen & Diamantopoulos, 2017), the original authors of the TIPI did not explicitly indicate whether the TIPI should be investigated reflectively or formatively. In fact, the point could be made that the original investigation of the TIPI assumes a reflective model (the items are caused by the construct), in that the statistical methods used to investigate it are reflective by nature (notably, Cronbach's α), and lower correlations between items are described as flaws that are imputable to the brevity of the measure (Gosling et al., 2003).

However, the point could also be made that the TIPI is constructed using Behavior Domain Theory (i.e., the items are samples of the construct) since the authors mention that their primary aim in selecting items is to cover the content of a construct broadly, rather than homogeneously. Indeed, as the original authors put it, "the TIPI instead emphasized content validity considerations, resulting in lower inter-item correlations than is typical of more homogenous scales" (Gosling et al., 2003, p. 516). The authors also explain that they first "strove for breadth of coverage" (Gosling et al., 2003, p. 508) and "aimed to enhance the bandwidth of the items by including in each item several descriptors selected to capture the breadth of the Big-Five dimensions" (Gosling et al., 2003, p. 508).

Although emphasizing content validity at the expense of internal consistency is a typical dilemma in the construction of short scales (Ziegler et al., 2014), this description of the objectives of the TIPI may signal that the initial objective of the authors was to representatively sample the traits. Thus, behavior domain theory may underlie the construction of the TIPI more than the Causal Theory of Measurement, and the discrepancy between the reflective methods used and the underlying framework of the test construction may explain the challenges encountered (e.g., lower inter-item correlations) in the investigations of the TIPI.

The two items of a personality trait thus represent a potentially different sample of a domain, and the items for each personality trait may not be necessarily correlated. If we reproduce the formative mental experiment (Bollen & Diamantopoulos, 2017), we find that, for example, it is perfectly conceivable that the two TIPI items of Agreeableness ("Critical, quarrelsome" and "Sympathetic, warm") may represent two attributes of (dis)agreeable individuals, while not necessarily having a common cause. In other words, we could imagine that an increase in Agreeableness would not necessarily imply that an individual is less critical-quarrelsome, but that a decrease in being critical-quarrelsome would impact one's Agreeableness.

1.5 | The aim of this study

The structure of the TIPI has been so far investigated with a reflective framework only, and mostly with limited success: We advance here that these results may not be a problem of the TIPI itself, but of its misconception and investigation as reflective. Indeed, we believe that the conjunction of (a) the low observed Cronbach's alpha values, (b) the poor fit of reflective measurement models to the TIPI's data, (c) the good observed convergent validity and test–retest reliability, (d) the purpose of the instrument to "cover content," and (e) the mental experiments (Bainter & Bollen, 2014; Bollen &

Diamantopoulos, 2017) on the items of the TIPI all point in one direction: The underlying measurement framework of the TIPI may be formative.

The aim of this research is thus to reinvestigate the TIPI using a psychometric procedure appropriate for a formative instrument: the formative MIMIC model (Diamantopoulos & Temme, 2013; Joreskog & Goldberger, 1975), which is more extensively discussed in the Method section.

We hypothesized that a formative MIMIC model would adequately fit the data collected through the TIPI and would show that the latent variables formed by the TIPI items are good predictors of the Big Five personality traits, measured using two different instruments: the Big Five Inventory (BFI; John & Srivastava, 1999) in Study 1 and the Revised NEO Personality Inventory (NEO-PI-R; Costa & McCrae, 1992b) in Study 2. The BFI and NEO-PI-R were selected as external validity criteria because they are among the most frequently used measures of the Big Five, which the TIPI claims to capture.

2 | STUDY 1

In Study 1, we investigated the TIPI using a formative MIMIC model, predicting BFI scores.

2.1 | Method

2.1.1 | Participants

A total of 938 French undergraduate students (579 females, 359 males; $M_{\text{age}} = 21.6$, $SD = 2.15$) in psychology and management volunteered to participate in the study.

2.1.2 | Procedure

The participants were successively administered the French TIPI and the BFI, described later. Both questionnaires and demographic questions were administered via computer. The participants responded anonymously and were told, prior to responding, that this study would consist of a general exploration of their personality traits. The participants were not compensated for responding.

2.1.3 | Instruments

Ten-Item Personality Inventory

The TIPI (Gosling et al., 2003) is a very short measure of the Big Five personality traits. It is composed of two items per personality dimension, with one of each pair being a reversed item. The participants responded using a 7-point Likert scale, indicating the extent to which they agree or disagree with each statement. As earlier explained, apart from its internal structure, the TIPI has shown remarkable psychometrical qualities for a scale of such extreme brevity, notably good test–retest reliability and satisfactory concurrent validity with various measures, including the NEO-PI-R (Gosling et al., 2003). We used its French translation (Storme et al., 2016), which has shown similar qualities.

Big Five Inventory

The BFI (John & Srivastava, 1999) is a 44-item inventory that aims at measuring the Big Five personality traits of the Five-Factor Model (Costa & McCrae, 1992a; McCrae & John, 1992): Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness. The participants responded using a 5-point Likert scale, indicating the extent to which they agreed or disagreed with each statement. Although not as short as the TIPI, the BFI is widely used as a short measure of the Big Five—it is actually especially used as a concurrent validity criterion in many psychometric investigations (e.g., Myszkowski, Storme, Zenasni, & Lubart, 2014)—and has been found to have satisfactory psychometric properties (John & Srivastava, 1999). Its French translation, which was used here, was notably found to have good internal consistency (Myszkowski & Storme, 2012; Plaisant, Srivastava, Mendelsohn, Debray, & John, 2005), as well as satisfactory convergent and discriminant validity when compared with the scores obtained with the NEO-PI-R (Plaisant et al., 2005).

2.1.4 | Statistical analyses

We used structural equation modeling (SEM) with maximum likelihood (ML) estimation, using the R package lavaan (Rosseel, 2012). The main hypothetical model that was fit to the data was a multidimensional formative MIMIC model (Bollen & Diamantopoulos, 2017; Diamantopoulos & Temme, 2013; Joreskog & Goldberger, 1975). In this model, five latent variables (corresponding to the Big Five personality traits) were each formed by their two corresponding items, and these five latent formative constructs predicted the five BFI scores. The BFI sum scores were used rather than their latent counterparts because, as is frequent for investigations of Big Five measures (Borkenau & Ostendorf, 1990; Marsh et al., 2010; Vassend & Skrondal, 1997), previous CFA investigations of the BFI have shown problematic fit (Leung, Wong, Chan, & Lam, 2012), and these issues could have lowered this study's model fit and thus made the model uninterpretable.

Although formative variables have by definition no residual variance (the model would not be identified), the fit of a formative MIMIC model has been argued to still be informative (Bollen & Diamantopoulos, 2017). More specifically, the fit of a formative MIMIC model can indicate (here, for each trait) whether there is a "single intervening latent variable" (Bollen & Diamantopoulos, 2017, p. 589)—as opposed to none or multiple intervening variables—in the relation

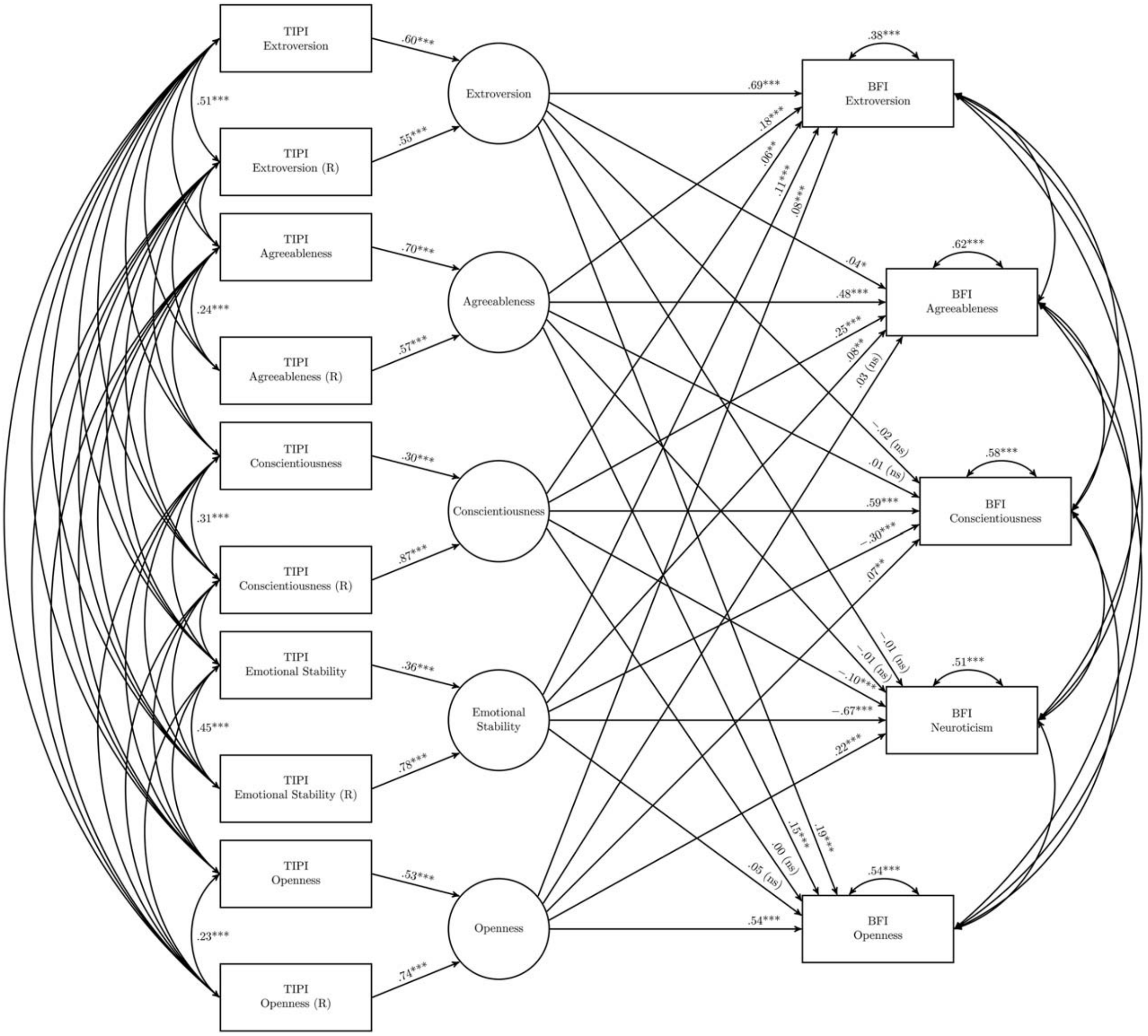

FIGURE 1 Path diagram of the TIPI–BFI formative MIMIC model with standardized estimates. *$p < .05$. **$p < .01$. ***$p < .001$

between the formative indicators (here, the TIPI items) and the outcome variables (here, the BFI scores). In other words, here, a good fit for a formative MIMIC model would suggest that, for each trait, the TIPI items predict the BFI scores totally (or at least, to a great extent) through the formative latent variable.

Multiple indices were used to assess the fit of the tested model to the data: the comparative fit index (CFI), the standardized root mean square residual (SRMR), and the root mean square error of approximation (RMSEA). The cut-off values for acceptable model fit used in this study were above .95 for the CFI (Hu & Bentler, 1999) and under .08 for the SRMR (Hu & Bentler, 1999) and the RMSEA (Browne & Cudeck, 1992).

2.2 | Results

The formative MIMIC model—reported in Figure 1, along with standardized estimates—had an acceptable fit to the data ($\chi^2 = 124.73$, $df = 20$, $p < .001$, CFI = .968, SRMR = .019, RMSEA = .075), allowing further interpretation of its estimates. Moderate to strong significant (all $|\beta|$s $> .48$, all ps $< .001$) effects between all five latent formative traits and their corresponding BFI scores were observed. The correlations between items of each TIPI scale ranged between .23 and .51, indicating that the homogeneity within these scales—a quality expected by reflective measurement models, but not by formative models—was not necessarily high.

The formative weights were significant for all items (all ps $< .001$) but revealed that the latent constructs were in some cases "better formed" by some items than others: The nonreversed Conscientiousness item notably had a weak ($\beta = .30$) formative weight on the Conscientiousness construct, whereas the reversed item had a much stronger weight ($\beta = .87$). It should, however, be noted that these weights are of course dependent upon the external criteria (here, the BFI traits) predicted by the latent variable.

Even though a reflective–formative model fit comparison is not possible here (Bollen & Diamantopoulos, 2017), we fit a model similar to the formative model presented in Figure 1, with the exception of the latent variables' being reflective. That model had an unsatisfactory fit ($\chi^2 = 1003.96$, $df = 65$, $p < .001$, CFI = .797, SRMR = .118, RMSEA = .124). As done in previous research on the TIPI (Oshio et al., 2014), we also fit reflective models that controlled for an overall shared method factor in the TIPI responses ($\chi^2 = 717.20$, $df = 56$, $p < .001$, CFI = .857, SRMR = .081, RMSEA = .112), a reversed items shared method factor ($\chi^2 = 932.7$, $df = 60$, $p < .001$, CFI = .811, SRMR = .111, RMSEA = .125), and both ($\chi^2 = 634.22$, $df = 52$, $p < .001$, CFI = .874, SRMR = .082, RMSEA = .109)—all of them yielding unsatisfactory fit.

Additionally, formative models (similar to Figure 1) without correlations between TIPI items ($\chi^2 = 1531.34$, $df = 65$, $p < .001$, CFI = .683, SRMR = .149, RMSEA = .155), without correlations between BFI scores ($\chi^2 = 255.10$, $df = 30$, $p < .001$, CFI = .931, SRMR = .026, RMSEA = .089), and without either ($\chi^2 = 1661.71$, $df = 75$, $p < .001$, CFI = .657, SRMR = .151, RMSEA = .150) were also tested, yielding unsatisfactory fit.

3 | STUDY 2

In Study 2, we investigated the TIPI using a formative MIMIC model, but this time predicting NEO-PI-R scores.

3.1 | Method

3.1.1 | Participants

A total of 759 French adults (474 females, 285 males; $M_{\text{age}} = 44.3$, $SD = 12.1$) from the general population volunteered to participate in the study.

3.1.2 | Procedure

The participants were successively administered the French TIPI and NEO-PI-R, described later. Both questionnaires and demographic questions were administered via computer. The participants responded anonymously and were told, prior to responding, that this study would consist of a general exploration of their personality traits. The participants were not compensated for responding.

3.1.3 | Instruments

Ten-Item Personality Inventory

The French TIPI (Storme et al., 2016) was administered, in a completely identical way as in Study 1.

NEO-PI-R

The NEO-PI-R (Costa & McCrae, 1992b) is one of the most researched self-report inventories. It measures each of the Big Five personality traits through six facets, and each facet is measured through through items, for a total of 240 items. The participants responded using a 7-point Likert scale. Although a more recent version of the NEO-PI exists (McCrae, Costa, & Martin, 2005), its French translation was not yet available at the time of the data collection. The French NEO-PI-R (Rolland, Parker, & Stumpf, 1998) has been shown to have psychometrical properties that are comparable with the original version.

3.1.4 | Statistical analyses

As in Study 1, we used structural equation modeling (SEM) with maximum likelihood (ML) estimation, using lavaan (Rosseel, 2012). We fit a multidimensional formative MIMIC model (Bollen & Diamantopoulos, 2017; Diamantopoulos & Temme, 2013; Joreskog & Goldberger, 1975), with five latent variables (corresponding to the Big Five personality traits) formed by their two corresponding TIPI items, and these five latent formative constructs predicted the five NEO-PI-R trait scores. As in Study 1, the NEO-PI-R sum scores were used rather than their latent counterparts because previous CFA investigations of the NEO-PI have shown problematic fit (Marsh et al., 2010; Vassend & Skrondal, 1997); these issues could have lowered this study's model fit and thus would have made the model uninterpretable. Similar to Study 1, multiple indices were used to assess the fit of the tested model to the data: the comparative fit index (CFI), the standardized root mean square residual (SRMR), and the root mean square error of approximation (RMSEA).

3.2 | Results

The tested model had an acceptable fit to the data ($\chi^2 = 116.60$, $df = 20$, $p < .001$, CFI = .941, SRMR = .025, RMSEA = .080), allowing further interpretation of its estimates.

The tested model is reported in Figure 2, along with standardized estimates. The standardized estimates showed significant moderate to strong effects (all $|\beta|$s $> .36$, all ps $< .001$) between all five latent formative traits and their corresponding NEO-PI-R scores. The correlations between items of each TIPI scale ranged between .21 and .41, indicating, similar to Study 1, a somewhat low homogeneity within the scales. Contrary to Study 1, the range of the formative weights was rather small—between .41 and .77—thus indicating that all items were somewhat comparably good indicators of their formative constructs.

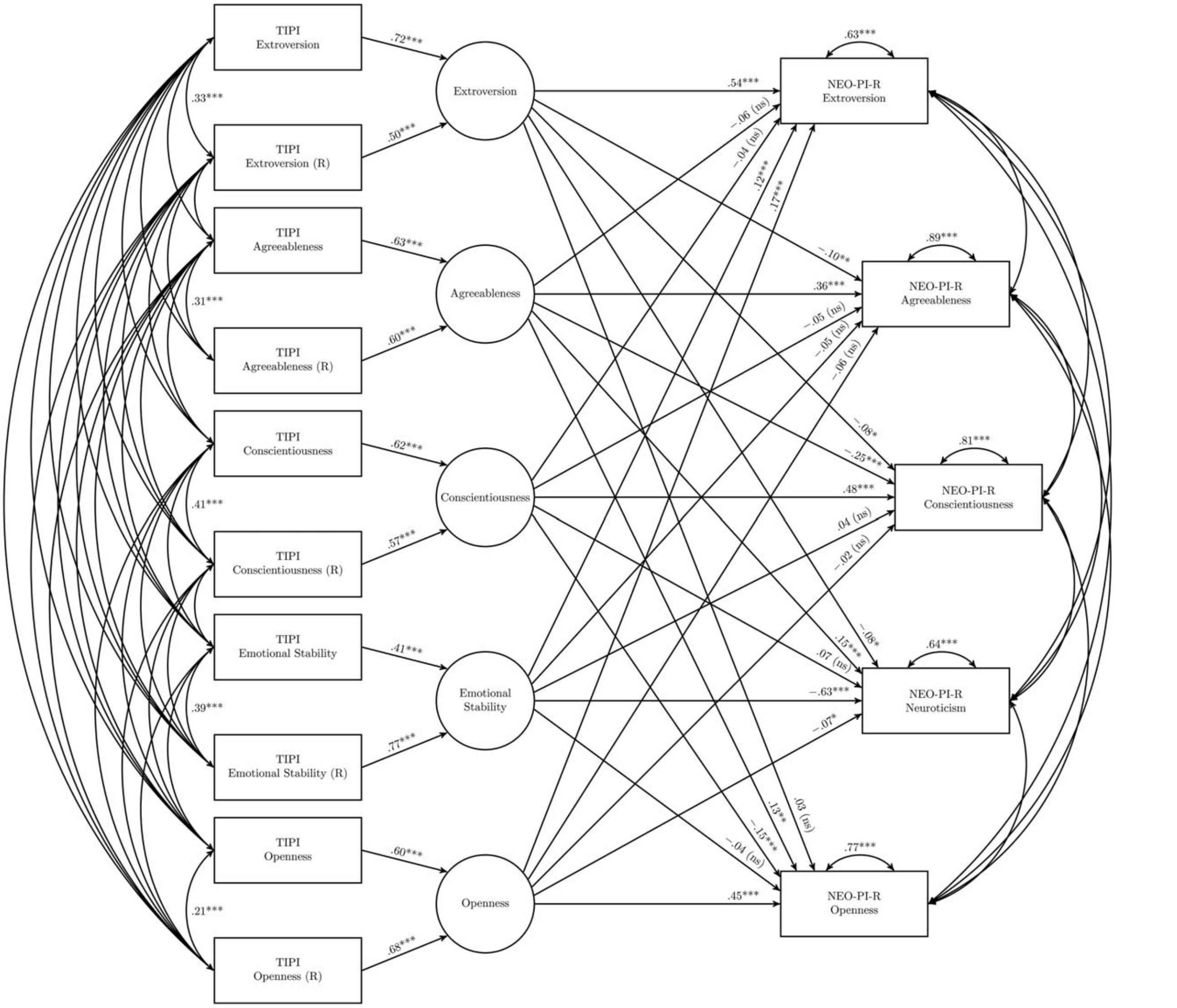

FIGURE 2 Path diagram of the TIPI–NEO-PI-R formative MIMIC model with standardized estimates. *$p < .05$. **$p < .01$. ***$p < .001$

Again, a reflective–formative model fit comparison is not possible here, but, for additional information, we fit a model similar to the formative model presented in Figure 2, with the exception of the latent variables' being reflective. That model had an unsatisfactory fit ($\chi^2 = 918.05$, $df = 65$, $p < .001$, CFI = .698, SRMR = .129, RMSEA = .131). As done in previous research on the TIPI (Oshio et al., 2014), we also fit reflective models that controlled for an overall shared method factor in the TIPI responses, which did not converge; a reversed items shared method factor ($\chi^2 = 827.36$, $df = 60$, $p < .001$, CFI = .729, SRMR = .118, RMSEA = .130); and both, which did not converge—all of them indicating unsatisfactory fit.

Additionally, a formative model (like Figure 2) with no correlations between TIPI items ($\chi^2 = 1353.76$, $df = 65$, $p < .001$, CFI = .544, SRMR = .143, RMSEA = .162), a formative model with no correlations between BFI scores ($\chi^2 = 558.16$, $df = 30$, $p < .001$, CFI = .677, SRMR = .058, RMSEA = .152), and both ($\chi^2 = 1795.31$, $df = 75$, $p < .001$, CFI = .392, SRMR = .153, RMSEA = .170) were also tested, indicating unsatisfactory fit.

4 | DISCUSSION

We argued that reflective models may not be in line with the TIPI's primary aim of representatively and broadly sampling the Big Five (Gosling et al., 2003). We thus suggested that its content actually implied behavior domain theory as an underlying measurement framework, and thus that its investigations may be formative.

This first investigation of the TIPI through a multidimensional formative MIMIC model (Bollen & Diamantopoulos, 2017; Diamantopoulos & Temme, 2013; Joreskog & Goldberger, 1975) has adequately fit the data in our samples, although we cannot compare the fit of the model tested in this study with the formative models because they have different dependent variables. Our results indicate that the TIPI items form latent constructs that mediate the relation between the items and other Big Five measures, here the BFI and the NEO-PI-R scores. We should note that these formative models do not assume a common cause to the items, but they can account for their being correlated, and here, the alternative models where item correlations were not included did not fit the data satisfactorily. Thus, it should not be concluded that the TIPI items are not correlated.

Previous investigations of the TIPI—whether its original version or its French translation used here—pointed to a specific lack of consistency for the Agreeableness and the Openness scales, suggesting their modification (Gosling et al., 2003; Storme et al., 2016). In contrast, our formative investigations would point out that the most problematic item (in terms of forming a latent variable that predicts an external criterion) could actually be the nonreversed Conscientiousness item, since this item had a weak weight in the Conscientiousness latent formative construct. This limitation was, however, not found in Study 2 with the NEO-PI-R: In this study, the range of the formative loadings was smaller, indicating less imbalance between the "formative power" of the different items. This discrepancy may indicate that the nonreversed Conscientiousness item better matches to the Conscientiousness construct measured by the NEO-PI-R than by the BFI (even though they theoretically are supposed to be the same). It may also be due to the differences in the psychometrical robustness of the criteria. Alternatively, since the two models are applied to different samples, this result may be due to sampling effects.

5 | LIMITATIONS

This study certainly has limitations. First, it investigates the French version of the TIPI only, on samples that have different biases, notably with one being a convenience sample of students. The results observed here call for replication in different samples, using different translations of the TIPI, and using other external criteria.

Second, this study is essentially a proof of concept only, since we do not demonstrate empirically that the formative approach is better. Instead, we argue that a formative investigation of short scales can be theoretically grounded and empirically feasible, through a concrete example. Nevertheless, although we do argue that formative investigations may be theoretically more relevant for some measures like the TIPI, and although we do show that formative investigations of the TIPI are at least empirically feasible, it is impossible (with current methods) to demonstrate empirically that they are better or more useful than reflective investigations, even in this specific case. The underlying measurement theory of an instrument and the structural model used to investigate it should certainly be aligned, but there is no way to empirically indicate which model or theory is correct here.

Third, formative models are a heated debate in measurement theory, and our research's being essentially a suggestion that formative models may be considered for some short scales, as well as the criticisms of formative models in general (for a list, see Bollen & Diamantopoulos, 2017), are logically applicable to this study.

Fourth, we suggested considering formative models for short scales because of their primary aim of breadth of content (Gosling et al., 2003; Ziegler et al., 2014), but a good fit when investigating them using a formative MIMIC model does not imply that the test development objective of adequately sampling a behavior domain is met. In other words, this study does not provide evidence that the content of the Big Five is sufficiently captured by the TIPI. Finding evidence that the formative constructs are good predictors of longer scale scores is certainly a good sign of a form of unidimensionality of the construct as a mediating variable (Bollen & Diamantopoulos, 2017), but other methods of validation may be used, such as content validity analyses.

6 | IMPLICATIONS FOR FUTURE RESEARCH

The results of this study call for a clarification of the approach to measurement used when building a scale (Markus & Borsboom, 2013). Indeed, reflective and formative measurement approaches are not only different in terms of the statistical tools used to investigate them, but they are also, more importantly, different in the approach to test construction: Is the objective to form a representative sample of a behavior domain or to measure the behavioral effects of a common cause? In the former case, broad content and formative investigations (formative MIMIC models with external criteria) are called for; in the latter case, reflective investigations (internal consistency, reflective latent variable models, etc.) are called for. Clarifying the item construction process, framework, and measurement goals leads to psychometric investigations that are more strategically targeted at testing whether this goal is achieved.

Beyond short scales, we think that this study certainly questions chasing breadth (representatively sampling the behavior in its domains) and consistency (observing causes of the same trait) at the same time in item construction. Psychological testing textbooks (e.g., DeVellis, 2016) often adopt a pragmatic, yet paradoxical, approach to this, explaining that one should develop items that are similar, but not too much. We may suggest that clearer processes of item construction and statements about the underlying framework be made. An example for a formative measure could be "the researchers attempted to capture the various domains in which that type of behavior may be observed," and an example for a reflective measure could be "the researchers attempted to find multiple instances in which the trait is expressed." Future research may clarify guidelines in reflective and formative test construction.

Developers of very short scales are "caught between a rock and a hard place" (Ziegler et al., 2014, p. 187), trying to achieve both internal consistency and content coverage. In other words, in a way, short scales change priorities in scale construction, as they often change the focus to breadth, leading to lower internal consistencies. A previously proposed solution to lower internal consistencies is to investigate very short scales using forms of reliability other than internal consistency, notably test–retest reliability (Gosling et al., 2003; Ziegler et al., 2014). Another proposed solution (Ziegler et al., 2014) is to use estimates of unidimensionality that are less biased by test length than Cronbach's alpha, such as McDonald's omega. As an alternative solution, we propose that when changing priorities by favoring breadth of content, the researchers may attempt to sample a behavior representatively rather than to find indicators of a common cause. Thus, their coverage of different domains would explain lower correlations between items of the same scale. Therefore, what we propose here is that the issue may not only reside in the flaws of Cronbach's alpha, but also in the absence of a clear decision of an underlying measurement framework, leading to psychometric investigations that may be inconsistent with the original test construction process and objectives.

ACKNOWLEDGMENT

The author(s) received no financial support for the research, authorship, and/or publication of this article.

CONFLICT OF INTERESTS

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

ORCID

Nils Myszkowski http://orcid.org/0000-0003-1322-0777

REFERENCES

Bainter, S. A., & Bollen, K. A. (2014). Interpretational confounding or confounded interpretations of causal indicators? *Measurement: Interdisciplinary Research and Perspectives*, *12*, 125–140. https://doi.org/10.1080/15366367.2014.968503

Bollen, K. A., & Diamantopoulos, A. (2017). In defense of causal-formative indicators: A minority report. *Psychological Methods*, *22*, 581–596. https://doi.org/10.1037/met0000056

Borkenau, P., & Ostendorf, F. (1990). Comparing exploratory and confirmatory factor analysis: A study on the 5-factor model of personality. *Personality and Individual Differences*, *11*, 515–524.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*, 230–258.

Cattell, R. B. (1943). The description of personality. I. Foundations of trait measurement. *Psychological Review*, *50*, 559–594. https://doi.org/10.1037/h0057276

Chiorri, C., Bracco, F., Piccinno, T., Modafferi, C., & Battini, V. (2014). Psychometric properties of a revised version of the Ten Item Personality Inventory. *European Journal of Psychological Assessment*, *31*, 109–119. https://doi.org/10.1027/1015-5759/a000215

Costa, P. T., & McCrae, R. R. (1992a). Four ways five factors are basic. *Personality and Individual Differences*, *13*, 653–665. https://doi.org/10.1016/0191-8869(92)90236-I

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

DeVellis, R. F. (2016). *Scale development: Theory and applications*. Thousand Oaks, CA: Sage.

Diamantopoulos, A., & Temme, D. (2013). MIMIC models, formative indicators and the joys of research. *AMS Review*, *3*, 160–170. https://doi.org/10.1007/s13162-013-0050-0

Diamantopoulos, A., & Winklhofer, H. M. (2001). Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research*, *38*, 269–277. https://doi.org/10.1509/jmkr.38.2.269.18845

Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, *59*, 1216–1229. https://doi.org/10.1037/0022-3514.59.6.1216

Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*, 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1

Hofmans, J., Kuppens, P., & Allik, J. (2008). Is short in length short in content? An examination of the domain representation of the Ten Item Personality Inventory scales in Dutch language. *Personality and Individual Differences*, *45*, 750–755. https://doi.org/10.1016/j.paid.2008.08.004

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1–55. https://doi.org/10.1080/10705519909540118

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 102–138). New York, NY: Guilford Press.

Joreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*, 631–639. https://doi.org/10.2307/2285946

Leung, D. Y., Wong, E. M., Chan, S. S., & Lam, T. H. (2012). Psychometric properties of the Big Five Inventory in a Chinese sample of smokers receiving cessation treatment: A validation study. *Journal of Nursing Education and Practice*, *3*, 1–10. https://doi.org/10.5430/jnep.v3n6p1

Markus, K. A., & Borsboom, D. (2013). Reflective measurement models, behavior domains, and common causes. *New Ideas in Psychology*, *31*, 54–64. https://doi.org/10.1016/j.newideapsych.2011.02.008

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J., Trautwein, U., & Nagengast, B. (2010). A new look at the Big Five factor structure through exploratory structural equation modeling. *Psychological Assessment*, *22*, 471–491. https://doi.org/10.1037/a0019227

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*, 261–270. https://doi.org/10.1207/s15327752jpa8403_05

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality*, *60*, 175–215.

Muck, P. M., Hell, B., & Gosling, S. D. (2007). Construct validation of a short five-factor model instrument: A self-peer study on the German adaptation of the Ten-Item Personality Inventory (TIPI-G). *European Journal of Psychological Assessment*, *23*, 166–175. https://doi.org/10.1027/1015-5759.23.3.166

Myszkowski, N., & Storme, M. (2012). How personality traits predict design-driven consumer choices. *Europe's Journal of Psychology*, *8*, 641–650. https://doi.org/10.5964/ejop.v8i4.523

Myszkowski, N., Storme, M., Zenasni, F., & Lubart, T. I. (2014). Appraising the duality of self-monitoring: Psychometric qualities of the Revised Self-Monitoring Scale and the Concern for Appropriateness Scale in French. *Canadian Journal of Behavioural Science/Revue Canadienne des Sciences du Comportement*, *46*, 387–396. https://doi.org/10.1037/a0033107

Oshio, A., Abe, S., Cutrone, P., & Gosling, S. D. (2014). Further validity of the Japanese version of the Ten Item Personality Inventory (TIPI-J). *Journal of Individual Differences*, *35*, 236–244. https://doi.org/10.1027/1614-0001/a000145

Plaisant, O., Srivastava, S., Mendelsohn, G. A., Debray, Q., & John, O. P. (2005). Relations entre le Big Five Inventory français et le manuel diagnostique des troubles mentaux dans un échantillon clinique français [Relations between the French version of the Big Five Inventory and the DSM classification in a French clinical sample of psychiatric disorders]. *Annales Médico-Psychologiques*, *163*, 161–167. https://doi.org/10.1016/j.amp.2005.02.002

Renau, V., Oberst, U., Gosling, S., Rusiñol, J., & Chamarro, A. (2013). Translation and validation of the Ten-Item Personality Inventory into Spanish and Catalan. *Aloma: Revista de Psicologia, Ciències de l'Educació i de l'Esport*, 31. Retrieved from http://www.revistaaloma.net/index.php/aloma/article/view/200

Rolland, J.-P., Parker, W. D., & Stumpf, H. (1998). A psychometric examination of the French translations of NEO-PI-R and NEO-FFI. *Journal of Personality Assessment*, *71*, 269–291. https://doi.org/10.1207/s15327752jpa7102_13

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36. https://doi.org/10.18637/jss.v048.i02

Storme, M., Tavani, J.-L., & Myszkowski, N. (2016). Psychometric properties of the French Ten-Item Personality Inventory (TIPI). *Journal of Individual Differences*, *37*, 81–87. https://doi.org/10.1027/1614-0001/a000204

Vassend, O., & Skrondal, A. (1997). Validation of the NEO Personality Inventory and the five-factor model. Can findings from exploratory and confirmatory factor analysis be reconciled? *European Journal of Personality*, *11*, 147–166. https://doi.org/10.1002/(SICI)1099-0984(199706)11:2<147::AID-PER278>3.0.CO;2-E

Ziegler, M., Kemper, C. J., & Kruyen, P. (2014). Short scales—Five misunderstandings and ways to overcome them. *Journal of Individual Differences*, *35*, 185–189. https://doi.org/10.1027/1614-0001/a000148

How to cite this article: Myszkowski N, Storme M, Tavani J-L. Are reflective models appropriate for very short scales? Proofs of concept of formative models using the Ten-Item Personality Inventory. *Journal of Personality*. 2019;87:363–372. https://doi.org/10.1111/jopy.12395