The Big Five Inventory–2

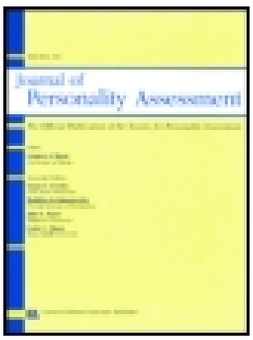
Denissen, Jaap J. A.; Geenen, Rinie; Soto, Christopher J.; John, Oliver P.; Van Aken, Marcel A. G.

Published in:
Journal of Personality Assessment

DOI:
10.1080/00223891.2018.1539004

Publication date:
2020

Document Version
Publisher's PDF, also known as Version of record



Journal of Personality Assessment

ISSN: 0022-3891 (Print) 1532-7752 (Online) Journal homepage: https://www.tandfonline.com/loi/hjpa20

The Big Five Inventory–2: Replication of Psychometric Properties in a Dutch Adaptation and First Evidence for the Discriminant Predictive Validity of the Facet Scales

Jaap J. A. Denissen, Rinie Geenen, Christopher J. Soto, Oliver P. John & Marcel A. G. van Aken

To cite this article: Jaap J. A. Denissen, Rinie Geenen, Christopher J. Soto, Oliver P. John & Marcel A. G. van Aken (2019): The Big Five Inventory–2: Replication of Psychometric Properties in a Dutch Adaptation and First Evidence for the Discriminant Predictive Validity of the Facet Scales, Journal of Personality Assessment, DOI: 10.1080/00223891.2018.1539004

To link to this article: https://doi.org/10.1080/00223891.2018.1539004

JOURNAL OF PERSONALITY ASSESSMENT
https://doi.org/10.1080/00223891.2018.1539004

The Big Five Inventory–2: Replication of Psychometric Properties in a Dutch Adaptation and First Evidence for the Discriminant Predictive Validity of the Facet Scales

Jaap J. A. Denissen[1], Rinie Geenen[2], Christopher J. Soto[3], Oliver P. John[4], and Marcel A. G. van Aken[2]

[1]Tilburg School of Social and Behavioral Sciences, Tilburg University, Tilburg, The Netherlands; [2]Department of Psychology, Utrecht University, Utrecht, The Netherlands; [3]Department of Psychology, Colby College; [4]Department of Psychology, University of California, Berkeley

ABSTRACT

This series of studies investigated whether the good psychometric properties of the English version of the Big Five Inventory–2 (BFI–2) could be replicated using its Dutch adaptation. Second, it aimed to further examine the predictive validity of both the Big Five domain and the more specific facet scales of the BFI–2 in a large and representative sample. Results indicated that the structure found in the English version was replicated in the Dutch adaptation. The 60-item BFI–2 was reliable at the level of both domains and facets, as were the abbreviated versions. In terms of validity, the domain scales predicted a broad range of criteria. Examination of preregistered hypotheses regarding the discriminant validity of the facets indicated that experts were able to predict which facets would be most strongly associated with specific criteria. Overall, results confirm the strong psychometric properties of the BFI–2 Big Five domain scales and indicate that theoretically identified facets can be more valid predictors of criteria than other facets of the same domain.

ARTICLE HISTORY
Received 4 July 2018
Revised 25 September 2018

The Big Five Inventory (BFI; see John & Srivastava, 1999) is a brief questionnaire using short phrases to measure the Big Five personality dimensions (extraversion, agreeableness, conscientiousness, negative emotionality, and open-mindedness). It is one of the most widely used questionnaires for personality assessment. Recently, a new version of the BFI, the BFI–2, was published, which has more balanced item formulations and includes three more specific facets within each broad domain[1] scale, thus promising more predictive precision. It is therefore likely that the BFI–2 will gain ground as a prominent instrument to measure the Big Five. To make a step forward in testing the validity of the broad, international application of BFI–2, the possibility of replication in different languages should be tested. A Dutch translation of the BFI was validated (Denissen, Geenen, Van Aken, Gosling, & Potter, 2008), but a validated Dutch BFI–2 version has not yet been developed. Moreover, although Soto and John (2017b) provided the first supportive evidence for the validity of the novel BFI–2 facets, this evidence was exploratory, and more confirmatory evidence is needed. These studies set out to achieve two main objectives: to evaluate the measurement properties of an adaptation of the BFI–2 in another language (Dutch) than the original English language, and to conduct confirmatory tests of the validity of the Big Five domain and the 15 facet scales included in the BFI–2.

Big Five trait dimensions: Structure, predictive validity, and facet structure

Structure and labeling of the Big Five

The Big Five make up a dimensional system of trait structure that captures five of the most important dimensions of variance underlying comprehensive trait ratings. Studies have investigated the replicability of the Big Five and found support for the generalizability of most of the dimensions across lexical studies in various cultural and linguistic settings (De Raad & Peabody, 2005). Despite the generalizability of the structure, the Big Five dimensions have been labeled differently across studies (for an overview, see Denissen & Penke, 2008). In this article, we adopt the nomenclature of Soto and John (2017b), who selected the labels extraversion, agreeableness, conscientiousness, negative emotionality, and open-mindedness.

Predictive validity and demographic correlates of the Big Five

Ozer and Benet-Martínez (2006) published a systematic overview of the predictive validity of the Big Five regarding individual, interpersonal, and social institutional outcomes. In their summary (see their Table 1), the following patterns

CONTACT Jaap J. A. Denissen jjadenissen@gmail.com Department of Developmental Psychology, Tilburg University, Tilburg, 5000 LE, The Netherlands.
This article was accepted under the editorship of Steven K. Huprich.

Supplemental data for this article can be accessed on the publisher's website.

[1]The term *domain* is used to refer to manifest individual differences corresponding to the broad Big Five dimensions.

Table 1. English and Dutch versions of the 60 Big Five Inventory–2 items and component loadings in Study 1.

Item no.	English version	Dutch version	Ext.	Agr.	Con.	Neg.	Ope.
	Extraversion	Extraversie					
	Sociability items	Gezelligheid items					
16	Tends to be quiet. (Original)	Doorgaans stil is	**−.78**	−.04	.05	−.03	.00
46	Is talkative. (Original)	Spraakzaam is	**.76**	.14	.03	.07	.14
1	Is outgoing, sociable. (Original)	Communicatief, een gezelschapsmens is	**.66**	**.34**	.05	.07	.03
31	Is sometimes shy, introverted. (Revised)	Soms verlegen, introvert is	**−.63**	.04	−.08	−.22	.06
	Assertiveness items	**Assertiviteit items**					
21	Is dominant, acts as a leader. (New)	De toon zet, als een leider handelt	**.52**	−.20	.15	.20	.29
6	Has an assertive personality. (Original)	Een persoon is die voor zichzelf opkomt	**.46**	−.07	.15	**.39**	.10
51	Prefers to have others take charge. (New)	Het liefst ziet dat anderen het voortouw nemen	**−.48**	.11	−.20	−.23	−.29
36	Finds it hard to influence people. (New)	Moeite heeft om andere mensen te overtuigen	**−.43**	.03	−.16	**−.31**	−−.25
	Energy level items	**Energie niveau items**					
41	Is full of energy. (Original)	Vol energie is	**.52**	.22	.22	**.33**	.06
56	Shows a lot of enthusiasm. (Revised)	Veel enthousiasme en uitbundigheid uitstraalt	**.69**	.26	.12	.08	.09
11	Rarely feels excited or eager. (New)	Zelden uitgelaten of gretig is	**−.50**	−.05	.05	.10	−.08
26	Is less active than other people. (New)	Minder levendig dan anderen is	**−.61**	−.21	−.03	−.20	−.17
	Agreeableness	Vriendelijkheid					
	Compassion items	**Compassie items**					
2	Is compassionate, has a soft heart. (New)	Betrokken, meevoelend is	.25	**.64**	.14	−.07	.12
47	Can be cold and uncaring. (Revised)	Koud en ongevoelig kan zijn	−.21	**−.63**	−.15	.08	−.07
32	Is helpful and unselfish with others. (Original)	Behulpzaam en onzelfzuchtig ten opzichte van anderen is	.12	**.47**	.19	.13	.16
17	Feels little sympathy for others. (New)	Weinig sympathie voor anderen voelt	−.11	**−.58**	−.10	.04	−.16
	Respectfulness items	**Respectvolle houding items**					
7	Is respectful, treats others with respect. (New)	Respectvol is, anderen met respect behandelt	.08	**.60**	.22	.12	.16
52	Is polite, courteous to others. (New)	Beleefd, hoffelijk tegenover anderen is	.05	**.62**	.23	.07	.11
37	Is sometimes rude to others. (Original)	Soms onbeleefd tegen anderen is	.01	**−.62**	−.22	−.03	−.03
22	Starts arguments with others. (Revised)	Snel ruzie maakt	.18	**−.52**	−.20	−.28	−.09
	Trust items	**Vertrouwen items**					
57	Assumes the best about people. (New)	Van het beste in mensen uitgaat	.14	**.62**	.00	.15	.07
27	Has a forgiving nature. (Original)	Vergevingsgezind en verdraagzaam is	−.03	**.64**	.00	.15	.09
12	Tends to find fault with others. (Original)	De neiging heeft om de fout bij anderen te zoeken	.01	**−.44**	−.14	−.16	−.09
42	Is suspicious of others' intentions. (New)	Niet zo snel uitgaat van de goede bedoelingen van anderen	−.02	**−.56**	.06	−.16	−.11
	Conscientiousness	Zorgvuldigheid					
	Organization items	**Organisatie items**					
3	Tends to be disorganized. (Original)	Geneigd is tot slordigheid	.06	−.02	**−.79**	.01	.04
18	Is systematic, likes to keep things in order. (New)	Systematisch is, dingen graag op orde houdt	.00	.05	**.72**	−.02	.10
33	Keeps things neat and tidy. (New)	Dingen netjes en verzorgd houdt	.09	.13	**.79**	.03	.02
48	Leaves a mess, doesn't clean up. (New)	Er een rommel van maakt, niet opruimt	−.02	−.16	**−.75**	−.06	−.02
	Productiveness items	**Productiviteit items**					
38	Is efficient, gets things done. (Revised)	Efficiënt is, klussen afkrijgt	.15	.09	**.57**	.29	.12
53	Is persistent, works until the task is finished. (Revised)	Volhoudend is, werkt tot de taak af is	.13	**.35**	**.52**	**.30**	.12
8	Tends to be lazy. (Original)	Geneigd is lui te zijn	−.17	−.23	**−.60**	−.14	−.01
23	Has difficulty getting started on tasks. (New)	Moeite heeft om met taken te beginnen	−.21	−.14	**−.52**	−.24	−.01
	Responsibility items	**Verantwoordelijkheid**					
28	Can be somewhat careless. (Original)	Een beetje nonchalant kan zijn *	.00	−.10	**−.59**	.12	.06
58	Sometimes behaves irresponsibly. (New)	Zich soms onverantwoordelijk en ondoordacht gedraagt	.00	**−.32**	**−.45**	−.16	−.09
43	Is reliable, can always be counted on. (Revised)	Betrouwbaar is, verwachtingen altijd waarmaakt	.14	**.32**	**.44**	.26	.16
13	Is dependable, steady. (New)	Verantwoordelijk, degelijk is	.11	**.34**	**.42**	.19	.13
	Negative emotionality	Negatieve emotionaliteit					
	Anxiety items	**Angst items**					
4	Is relaxed, handles stress well. (Original)	Ontspannen is, goed met stress kan omgaan	.14	.09	.06	**.74**	.06
34	Worries a lot. (Original)	Zich veel zorgen maakt	−.13	−.01	.08	**−.69**	.04
49	Rarely feels anxious or afraid. (New)	Zich zelden angstig of bang voelt	.09	.06	−.07	**.64**	−.01

19	Can be tense. (Original)	Gespannen kan zijn	−.08	−.01	−.04	**−.63**	.06
	Depression items	**Depressie items**					
39	Often feels sad. (New)	Zich vaak verdrietig voelt	−.20	−.06	−.07	**−.68**	.01
54	Tends to feel depressed, blue. (Revised)	Ertoe neigt zich terneergeslagen, somber te voelen	−.26	−.15	−.09	**−.70**	−.03
24	Feels secure, comfortable with self. (New)	Zich zeker, op zijn gemak met zichzelf voelt	**.31**	.12	.16	**.51**	−.01
9	Stays optimistic after experiencing a setback. (New)	Optimistisch blijft na een tegenslag	.21	.24	−.01	**.56**	.15
	Emotional volatility items	**Emotionele wisselvalligheid items**					
29	Is emotionally stable, not easily upset. (Original)	Emotioneel stabiel is, niet gemakkelijk overstuur	.06	.05	.13	**.76**	.07
59	Is temperamental, gets emotional easily. (New)	Opvliegend is, makkelijk emotioneel wordt	.17	−.21	−.17	**−.59**	−.08
44	Keeps their emotions under control. (New)	Zijn/haar emoties onder controle houdt	−.13	.07	.20	**.60**	.04
14	Is moody, has up and down mood swings. (Revised)	Humeurig is, wiens stemming op en neer gaat	−.06	**−.30**	−.15	**−.55**	−.03
	Open-mindedness	Ruimdenkendheid					
	Intellectual curiosity items	**Intellectuele nieuwsgierigheid items**					
55	Has little interest in abstract ideas. (New)	Weinig interesse in abstracte ideeën heeft	−.04	−.13	−.02	−.03	**−.64**
40	Is complex, a deep thinker. (Revised)	Genuanceerd en diep over dingen nadenkt	−.05	.18	.14	−.03	**.52**
25	Avoids intellectual, philosophical discussions. (New)	Intellectuele, filosofische discussies uit de weg gaat	−.15	.03	.02	−.16	**−.63**
10	Is curious about many different things. (Revised)	Benieuwd is naar veel verschillende dingen	.22	.20	.06	.13	**.51**
	Aesthetic sensitivity items	**Esthetische gevoeligheid items**					
20	Is fascinated by art, music, or literature. (Revised)	Gefascineerd is door kunst, muziek of literatuur	−.04	.10	−.06	−.10	**.69**
5	Has few artistic interests. (Original)	Weinig interesse voor kunst heeft	.07	−.07	.10	.07	**−.73**
35	Values art and beauty. (Revised)	Waarde hecht aan kunst en schoonheid	−.03	.13	.02	−.04	**.65**
50	Thinks poetry and plays are boring. (New)	Vindt dat dichtkunst en toneel maar saai zijn	−.01	−.18	.04	.10	**−.57**
	Creative imagination items	**Creatieve verbeelding items**					
30	Has little creativity. (New)	Weinig creativiteit heeft	−.22	−.12	−.10	−.07	**−.57**
15	Is inventive, finds clever ways to do things. (Revised)	Vindingrijk is, creatieve manieren verzint om dingen te doen	.18	.08	.14	.17	**.53**
60	Is original, comes up with new ideas. (Original)	Origineel is, met nieuwe ideeën komt	**.33**	.04	.10	.17	**.55**
45	Has difficulty imagining things. (New)	Weinig verbeeldingskracht heeft	−.24	−.09	−.09	−.02	**−.57**

Table 2. Coefficient alpha for the scales of the full (60 items), short BFI–2S (30 items), and extra-short (15 items) versions of the Big Five Inventory 2 BFI–2.

	Full Domain	Facet	Short Domain	Facet	Extra-short Domain
Extraversion	.86		.77		.62
Sociability		.80		.66	
Assertiveness		.72		.64	
Energy level		.70		.62	
Agreeableness	.84		.72		.56
Compassion		.68		.53	
Respectfulness		.68		.45	
Trust		.67		.40	
Conscientiousness	.87		.75		.64
Organization		.84		.74	
Productiveness		.74		.54	
Responsibility		.55		.21	
Negative emotionality	.88		.80		.72
Anxiety		.75		.59	
Depression		.76		.58	
Emotional volatility		.72		.58	
Open-mindedness	.85		.73		.66
Intellectual curiosity		.67		.45	
Aesthetic sensitivity		.83		.68	
Creative imagination		.79		.68	
M	.86	.73	.76	.58	.64

Note. Average values after Fisher transformation and back-transformation.

of validity evidence were reported. For extraversion, high levels have been associated with positive individual outcomes, such as life satisfaction and psychological and physical health. The trait has also been associated with higher relationship quality and satisfaction, and with higher levels of occupational and institutional involvement. For agreeableness, high levels have been associated with a mix of positive and negative outcomes that are consistent with this trait's conceptualization as a higher prioritization of the outcomes of others versus the self. Specifically, the trait has been positively associated with forgiveness, relationship satisfaction, and reduced crime. For conscientiousness, high levels have been consistently associated with positive health behaviors and outcomes, and also with occupational success. The trait negative emotionality has been negatively associated with several aspects of well-being, and positively associated with psychopathology. Finally, high levels of open-mindedness have been associated with spirituality and substance use, and also with more liberal political attitudes and values. In this article, we used the literature review by Ozer and Benet-Martínez (2006) to select criterion variables from an extensive panel study to validate the BFI–2's domain and facet scales.

Demographic correlates are an important part of the nomological network of the Big Five personality traits. These correlates can be used to validate novel personality scales, as was done by Denissen, Geenen, Selfhout, and van Aken (2008) in the case of the Dutch BFI translation. Meta-analytic evidence has suggested the following patterns. Regarding age correlates, research has established age-related increases in agreeableness and conscientiousness, and decreases in negative emotionality across adulthood (Roberts, Walton, & Viechtbauer, 2006; Srivastava, John, Gosling, & Potter, 2003; see also Lehmann, Denissen, Allemand, & Penke, 2013; Soto, John, Gosling, & Potter, 2011). Evidence for extraversion and open-mindedness is less consistent and has not indicated a

clear-cut linear trend at the level of the broad Big Five domains (Roberts et al., 2006). Regarding gender differences, a meta-analysis by Feingold (1994) found that men have lower levels of traits related to negative emotionality and agreeableness than women. A subsequent cross-cultural study was consistent with this analysis and found that men reported lower levels of negative emotionality and agreeableness, in addition to lower levels of extraversion and conscientiousness (Schmitt, Realo, Voracek, & Allik, 2008). Finally, results by Soto and John (2017b) using the BFI–2 were mostly consistent with these previous findings, although with notably attenuated gender differences in negative emotionality.

Predictive validity and demographic correlates of Big Five facets

Facets are more specific trait constructs located below the Big Five domain level. The BFI–2 has relied on theoretical considerations and the empirical literature to identify and define the three most prominent facets for each domain, but has subsequently refined these facets empirically using structural analyses to maximize convergent and discriminant relations, both at the domain and at the facet level (Soto & John, 2017b). The predictive validity of facets has been a source of debate. Already in 1957, Cronbach and Gleser (1957) described the so-called bandwidth-fidelity trade-off, that broader constructs (e.g., the Big Five) might predict a wider range of criteria, but more narrow constructs (e.g., facets) might be better suited to predict specific criteria.

Some findings indeed suggest that facets are able to predict incremental variance, over and above the Big Five domains. For example, Hagger-Johnson and Whiteman (2007) reported that the conscientiousness facet of self-discipline predicted aggregated health behaviors, over and above the conscientiousness domain scale. Similarly, MacCann, Duckworth, and Roberts (2009) showed that high school students' SAT test scores were better predicted by a conscientiousness facet they called perfectionism than by conscientiousness in general, and (low) absenteeism was better predicted by industriousness. Furthermore, Klimstra, Luyckx, Hale, and Goossens (2014) focused on longitudinal predictors of externalizing behavior and found that associations within certain domains were facet-specific (e.g., the reported association between extraversion and alcohol abuse was mainly due to the facets of sociability and activity). In another study, Klimstra, Luyckx, Goossens, Teppers, and De Fruyt (2013) found that associations between negative emotionality and ruminative exploration of identity were mainly due to internalizing facets of negative emotionality (e.g., anxiety, depression, etc.). Mund and Neyer (2014) also reported facet-specific effects, such as the finding that the conscientiousness facet of dependability (but not the total conscientiousness domain score) predicted decreases in insecurity in relationships with kin, and the extraversion facet of activity (but not the total extraversion domain score) predicted increases in closeness in relationships with kin. Finally, Terracciano et al. (2009) found that low scores on the Big Five conscientiousness domain scale predicted being overweight, but the strongest predictive associations were found for the facets of impulsiveness and order.

Some studies have also looked at the discriminant validity of the facets in terms of demographic criteria. For example, Weisberg, DeYoung, and Hirsh (2011) compared facets between men and women and found that women scored higher than men on a facet called orderliness—even though no gender difference in the overarching trait of conscientiousness was found. In a cross-cultural study by McCrae, Terracciano, and the members of the Personality Profiles of Cultures Project (2005), women were found to score higher on dutifulness and order but relatively lower on (self-endorsed) competence. Also, women were found to score higher on the extraversion facet of enthusiasm, whereas men scored higher on dominance—thus masking gender differences in the overall trait score. This is consistent with McCrae et al. (2005), who found that women scored higher on warmth, gregariousness, and positive emotions but relatively lower on assertiveness and excitement seeking. For open-mindedness, Weisberg et al. (2011) reported that women scored higher than men on the aesthetic aspects (i.e., enjoyment of art, beauty, and fantasy), whereas men scored higher on intellectual aspects (i.e., enjoyment of effortful thinking). McCrae et al. (2005) reported relatively consistent patterns, with women scoring higher on aesthetics, feelings, and actions but relatively lower on openness to ideas. Also consistent with this, Soto and John (2017b) found that women scored higher on aesthetics and men higher on intellectual curiosity facets of open-mindedness, but no overall differences in openness were found.

In a similar fashion, differences between facets have been reported regarding age correlates. Roberts et al. (2006) found that age trends differed between two facets of extraversion: Whereas social dominance increased with age, levels of social vitality neither increased nor decreased. Soto et al. (2011) found differences between facets in terms of age correlates, especially for extraversion, conscientiousness, and negative emotionality. For example, the conscientiousness facet of self-discipline showed a much sharper decrease during adolescence and a much sharper increase during adulthood when compared to the conscientiousness facet of order (which showed a similar pattern, but to a lesser extent).

Overall, scattered evidence thus seems consistent with the hypothesis that facets have incremental validity in addition to Big Five domains. That said, it has been argued that the incremental predictive validity correlations of facets that have emerged in previous research might be due to chance (Ones & Viswesvaran, 1996). Indeed, with the exception of facet-specific gender differences (see earlier), few of the facet-specific effects reviewed here have been closely replicated. It is possible to correct for chance capitalization but this would require huge sample sizes because facets, by definition, belong to the same domain and are thus substantially intercorrelated. A better solution is therefore a confirmatory hypothesis testing approach: Before any analyses are performed, specific hypotheses are derived about facets that demonstrate incremental predictive validity

compared to other facets for any specific validity criterion; the corresponding hypotheses can then be preregistered. To our knowledge, this research is the first to test preregistered hypotheses regarding the discriminant validity of Big Five facets.

The Big Five Inventory–2: Development and psychometric properties

More than 20 years after the BFI was created, Soto and John (2017b) developed a second, revised version of the BFI. This version introduced a number of innovations. First of all, the items of some BFI scales (primarily open-mindedness) were not balanced in terms of true-keyed and false-keyed items, thus making it more difficult to differentiate valid responses from more general response biases, such as acquiescence (the tendency for an individual to consistently agree or consistently disagree with survey items, regardless of their content). Therefore, the BFI–2 has an equal number of positive and negative items for each of the 5 domain and the 15 facet scales. Second, and perhaps more important, the BFI–2 consists of three prespecified facets per Big Five domain, with one facet being a "pure" or core manifestation of the underlying dimension (e.g., organization for conscientiousness), and the other two facets being theoretically meaningful variations, consistent with previous research and theorizing (e.g., productiveness and responsibility; DeYoung, Quilty, & Peterson, 2007; Saucier & Ostendorf, 1999; see Table 1 in Soto & John, 2017b, p. 121). The only exception to this principle is open-mindedness, where Soto and John (2017b) did not identify a core facet but instead regarded all facets as potentially equally important.

Soto and John (2017b) investigated the psychometric properties of the English-language BFI–2, using U.S. samples. They found evidence for the reliability, structure, and convergent and predictive validity of the instrument. Cronbach's alpha of the domain scales exceeded .80 across two studies, and the average Cronbach's alpha of the facet scales was higher than .75 across two samples, ranging between .66 and .85. Also, the test–retest stability of the domain scales was at least .76, and the average 2-month stability of the facet scales was .73 (range = .66–.83). The BFI–2 demonstrated clear evidence of factorial validity (e.g., high primary loadings compared to secondary loadings), and the distinctive nature of the facets was confirmed in a series of confirmatory factor analyses. The BFI–2 domain scales also converged substantially with corresponding scales of other Big Five instruments, such as the NEO PI–R (Costa & McCrae, 1992) and the Big Five Aspect Scales (DeYoung et al., 2007). Finally, the predictive validity of the facets was demonstrated by means of associations with a set of self- and peer-reported criteria. Moreover, when entered as a regression block, the BFI–2 facets provided a substantial degree of incremental predictive power over and above the variance predicted by the domain scales. In many cases, only one of the three facets in a domain seemed to predict individual criteria (e.g., only the conscientiousness facet of productiveness, but not responsibility or organization, predicted self-reported school achievement).

In spite of this impressive support, some limitations in the Soto and John (2017b) paper must be acknowledged. First, the BFI–2 scales were only validated against self-reports of well-being and behaviors associated with values, as well as peer reports of social connectedness, likability, stress resistance, and positive affect. Although this is impressive, these criteria are not broadly representative of outcomes that have been shown to be predicted by personality traits (Ozer & Benet-Martínez, 2006), thus potentially painting an incomplete picture. Second, although Soto and John (2017b) used a community sample to generate item content for the construction of the new BFI–2, the resulting instrument was validated in convenience samples of Internet users (Study 2) and university students (Study 3). It remains to be seen whether the good psychometric properties of the instrument also apply in a more heterogeneous and representative sample of the general population.

Finally, the discriminant validity of the facets was only tested in an exploratory fashion, which created the possibility for chance capitalization (Ones & Viswesvaran, 1996). Specifically, Soto and John (2017b) used hierarchical regression analysis to demonstrate the predictive power of the 5 BFI–2 domain scales and the 15 facet scales (both entered as a block; see Paunonen & Ashton, 2001, for a similar approach). They demonstrated that the BFI–2 was only slightly more predictive than the BFI when using the broad domain scales. However, when the domain scales of the BFI–2 were compared with the BFI–2 facet scales, the predictive validity of the block of 15 facets was significantly larger than a block of only the 5 domain scales (mean percentage of explained variance of 33% vs. 27%, respectively). However, these results only supported the validity of the facets as a block; they remained silent regarding the construct validity of any particular BFI–2 facet. To demonstrate this latter feature, a priori hypotheses regarding the pattern of predictive validity associated with each facet must be specified and tested.

This research

This research was designed with two key goals in mind. First, we wanted to examine whether the good psychometric properties of the original BFI–2 would replicate in samples of Dutch adults. After translating and backtranslating the BFI–2 items, we collected a series of convenience samples and used the corresponding principal component solutions to calibrate our item formulations (see supplementary materials for a description). Following this, we collected a large and representative sample to test a final version and establish the convergence of the BFI–2 domain scales with the scales of alternative Big Five instruments. Our second goal was to test the predictive validity of both the domains and the facets of the BFI–2 using a broad range of criteria. A unique feature of our study was that we preregistered hypotheses regarding the expected unique predictive validity of each assessed personality facet. We pursued these goals

through two studies. Specifically to address the first goal of the article, Study 1 was conducted to establish the factorial validity, internal consistency, and convergent validity of the Dutch BFI–2. To address the second goal, Study 2 was conducted to test preregistered hypotheses about the predictive validity of the domain and facet scales.

Study 1

Study 1 addressed the first goal of our article: to evaluate the measurement properties of the BFI–2 in another language and culture. We divided this goal into two subgoals. First, we wanted to provide a test of the Dutch BFI–2 in terms of structure and reliability. We started with exploratory principal component analyses and then ran confirmatory factor analyses to scrutinize the BFI–2 facet structure in more detail. Our second subgoal was to provide evidence for the convergent validity of the BFI–2 domain scales, by examining their associations with the frequently used International Personality Item Pool (IPIP) Big Five scales (Goldberg et al., 2006).

Method

Procedure

Participants of Study 1 were part of the Longitudinal Internet Studies for the Social Sciences (LISS) panel, which included monthly questionnaires (for details, see Scherpenzeel, 2011; Scherpenzeel & Das, 2010). The BFI–2 was added to the wave of data collection in July 2017.

Sample

Since 2008, LISS has followed a representative sample of the Dutch population based on a random sample of households drawn from the population register. An agreement with the LISS study team was made to draw at least 800 participants from the total panel (consisting of 7,000 individuals), and invitations were sent to 1,135 panel members. In total, 827 individuals (73%) completed the BFI–2, slightly exceeding the target. Missing data analysis using the R package MissMech (Jamshidian, Jalal, & Jansen, 2014; Version 1.0.2) indicated that missingness was not completely at random as indicated by their scores on the IPIP scales, which had been obtained in an earlier round of data collection from this sample, $p < .05$. Post-hoc tests showed that nonresponders were slightly less conscientious ($d = -.24$) and higher on neuroticism ($d = .19$). We had chosen to sample an equal number of males and females, and also an equal number across age groups. This succeeded, because 411 (50%) were female, and the sample was relatively balanced regarding age, with a slight overrepresentation of older individuals. Specifically, 135 (16%) participants were between 18 and 30 years old, 148 (18%) were between 31 and 43, 171 (21%) were between 44 and 56, 198 (24%) were between 57 and 69, and 172 (21%) were between 70 and 83.

Materials

In two preliminary studies (see supplementary materials), we created a well-functioning set of 56 translated BFI–2 items, with only 4 items of the 60-item set still being slightly problematic. During the second preliminary study, we pilot tested an alternative translation for a problematic low agreeableness item (Item 22, "Starts arguments with others," which had only weak loadings on agreeableness but strong loadings on extraversion), using 64 participants. Because this alternative version was already part of the BFI, where it had performed very well, we felt confident that this item would also perform well in our final instrument. To maximize the likelihood of ending up with satisfactory alternatives for each of the three remaining problematic items, we created potential alternatives for the following items: 36, "Finds it hard to influence people" (reverse-keyed item from the assertiveness facet of extraversion; one potential alternative), 42, "Is suspicious of others' intentions" (trust facet of agreeableness; two potential alternatives), and 43, "Is reliable, can always be counted on" (responsibility facet of conscientiousness; two potential alternatives). The BFI–2 items were administered (see Table 1), supplemented with the five potential alternative items (see supplementary materials, Table S3).

Two or 3 months prior to the BFI–2 assessment, 740 participants had completed the IPIP (Goldberg et al., 2006) Big Five scales. The resulting scores were used to estimate the convergent validity of the BFI–2 domains. Each of the IPIP Big Five scales consists of 10 items, such as "Am the life of the party" (extraversion) or "Spend time reflecting on things" (intellect). Cronbach's alphas of the scales ranged between .77 (intellect) and .89 (neuroticism) in the LISS sample.

Results

Deciding between parallel item versions

As a first step, the 65 items (including 5 parallel items) were analyzed in a principal component analysis after within-person centering each participant's item responses around their mean response to the complete set of items, to control individual differences in acquiescent responding (cf. Soto & John, 2017b). (This correction was only used for the exploratory principal component analysis.) In this analysis, we only focused on the performance of the parallel items on the basis of the component loading matrix. We selected the item versions with the highest primary loading and the largest differences between primary and secondary loadings (see Table S3 for details), resulting in our final set of 60 items.

Reliability analysis

Based on the 60-item set, we computed Cronbach's alpha coefficients as an indicator of internal consistency and present the results in Table 2. On average, the Cronbach's alphas of the domain scales were good, with coefficients averaging .86 and ranging from .84 (agreeableness) to .87 (conscientiousness). The Cronbach's alphas of the facet

scales were generally adequate, with an average of .73. Specifically, Cronbach's alphas were close to .70 or higher in all cases, except the responsibility facet of conscientiousness, which demonstrated an alpha of .55. A closer inspection indicated that this was due to Item 28 ("Can be somewhat careless"). We return to this issue in the discussion.

Besides the 60-item version, Soto and John (2017a) published short (30-item) and extra-short (15-item) versions of the BFI–2, with acceptable to good reliabilities for short Big Five domain scales. Table 2 also displays the Cronbach's alphas of these shorter scales in the Dutch adaptation. As can be seen, the reliability of the short scales was still satisfactory at the domain level (six items per scale), but the reliability of the short two-item facet scales dropped below the .60 threshold in many instances. Regarding the three-item extra-short Big Five domain scales published by Soto and John (2017a), Cronbach's alphas were even lower, as expected due to their limited length, and dropped below .60 in the case of agreeableness.

Discriminant correlations between facets and domain scales

To provide an initial overview of the internal structure of the BFI–2, we computed correlations between the facets and domain scales across all Big Five dimensions. As can be seen in Table S4, all but one correlation between the facets belonging to the same Big Five domain (marked by boxes) were substantial in size (around .50 or higher). There were some correlations higher than .30 between domains and facets of noncorresponding Big Five domains, and two correlations between domains were higher than .40: between extraversion and negative emotionality, and between agreeableness and conscientiousness. We return to this issue in the discussion.

Exploratory principal component analysis of domain-level structure

The within-person centered 60-item set was analyzed using principal component analysis followed by Varimax rotation. The scree plot indicated a clear five-component solution, in that the eigenvalues of the first 10 unrotated components were 12.05, 4.67, 4.44, 3.21, 2.75, followed by 1.73, 1.44, 1.23, 1.11, and 1.05. As shown in Table 1, the primary loadings of the items on their targeted components were substantial and averaged .60 (range = .42–.79). The average of each item's highest secondary loading was only .21 (range = .05–.39). The Dutch loadings were on average only .02 smaller than the English loadings obtained by Soto and John (2017b; see supplementary materials for a comparison). The patterns of the component loadings were also very similar between the Dutch loadings and the average loadings across the two validation samples of Soto and John (2017b)'s Study 3, with congruence coefficients ranging between .95 and .97.

Following Soto and John (2017b), we created facet scale scores and then used these as input for an exploratory principal component analysis, using Varimax rotation and fixing the number of components to five. The resulting loadings are presented in Table S5. As can be seen, primary loadings were very high, with a mean of .79. Secondary loadings were generally small and those reaching .30 or above (shown in bold) were consistent with expectations and prior findings. For example, given the link to low energy, the depression facet of negative emotionality had a notable negative loading on extraversion (which includes energy level as a facet). Similarly, the respectfulness facet of agreeableness (which involves following rules and standards but in the interpersonal domain) had a positive secondary loading on conscientiousness. Again, this pattern of primary and secondary loadings was very similar to the English component structure published by Soto and John (2017b, see Table 7, p. 132), with congruence coefficients ranging from .90 to .98.

Confirmatory factor analysis of facet-level structure

Soto and John (2017b) compared several structural models within each Big Five domain and found that the intended three-facet structure plus an acquiescence factor had the best fit in all cases. In this study, we tried to replicate these results with the Dutch BFI–2, comparing a baseline and an acquiescence-corrected facet-level model. In the baseline model, we defined a general domain factor, with loadings of all 12 items from each domain. In the facet-level models, we specified a structural matrix that allowed a loading of each item on its designated facet (and a loading of 0 on the two nondesignated facets), as well as an acquiescence factor for that domain (with all 12 item loadings fixed to 1). All models converged without problems, except the acquiescence-corrected facet model for open-mindedness, which triggered an error about a negative variance for Facet 1. We subsequently fixed this variance to a very small number (0.001) to address the issue.

Overall results of these confirmatory analyses are listed in Table S.6. As can be seen, the fit of the baseline models was poor, replicating the results of Soto and John (2017b). By comparison, the fit of the three facet-level models including an acquiescence method factor was acceptable to good in all cases, with all comparative fit index (CFI) values at least 0.92, all Tucker–Lewis Index (TLI) values at least 0.89, and all root mean square error of approximation (RMSEA) values at or below the benchmark of .08.

Convergent validity

We correlated the BFI–2 domain and facet scales with the five IPIP scales, which differ somewhat in label and definition, especially for openness, which is defined more narrowly as intellect. As can be seen in Table 3, convergent correlations of the domain scales ranged between .61 and .81, with an average of .72. These values are comparable to previous findings regarding convergence between the BFI and lexical measures of the Big Five (e.g., John, Naumann, & Soto, 2008; Soto & John, 2017b). Regarding the facets, convergent correlations with the IPIP scales were generally high as well, with only three moderate correlations ($.30 < r < .50$) being observed: Not surprisingly, IPIP intellect correlated only .35 with the BFI–2 openness facet of aesthetic

Table 3. Convergent and divergent correlations between Big Five Inventory–2 (BFI–2) domain and facet scales and the corresponding IPIP domain scales (no facets).

	BFI–2 domain		BFI–2 Facet 1		BFI–2 Facet 2		BFI–2 Facet 3	
IPIP domain scales (alternative BFI–2 label)	Con	Div	Con	Div	Con	Div	Con	Div
Extraversion	.76	.29	.76	.19	.53	.27	.60	.27
Agreeableness	.61	.25	.66	.22	.44	.22	.45	.20
Conscientiousness	.73	.21	.68	.11	.60	.23	.57	.21
Neuroticism (negative emotionality)	.81	.19	.74	.12	.68	.24	.68	.16
Intellect (open-mindedness)	.62	.21	.61	.20	.35	.10	.58	.22

Note. IPIP = International Personality Item Pool Big Five scale (10 items per scale); con = convergent correlation; div = divergent correlation (average correlation after reversing all scores related to negative emotionality). Note that the IPIP labels and definitions of the five domains are slightly different from those of the BFI–2 (see alternative BFI–2 label in brackets where relevant). For BFI–2 extraversion, Facet 1 = sociability, Facet 2 = assertiveness, Facet 3 = energy level. For BFI–2 agreeableness, Facet 1 = compassion, Facet 2 = respectfulness, Facet 3 = trust. For BFI–2 conscientiousness, Facet 1 = organization, Facet 2 = productiveness, Facet 3 = responsibility. For BFI–2 negative emotionality, Facet 1 = anxiety, Facet 2 = depression, and Facet 3 = emotional volatility. For BFI–2 open-mindedness, Facet 1 = intellectual curiosity, Facet 2 = aesthetic sensitivity, and Facet 3 = creative imagination. Bivariate correlations of $\geq .10$ would be statistically significant at $p < .05$ if tested.

sensitivity; IPIP agreeableness correlated .44 and .45 with respectfulness and trust. In all cases, the first facet, which should theoretically capture the core of each domain (except for open-mindedness), correlated most highly with the corresponding IPIP scale.

Discussion

A first goal of Study 1 was to check whether the Dutch BFI–2 items would replicate the measure's intended five-dimensional structure. Overall, this was clearly the case: The pattern of primary and secondary component loadings corresponded closely to the one reported by Soto and John (2017b). Specifically, using exploratory principal component analysis, our items had large primary loadings, and only minor secondary loadings. In the few cases where a secondary loading was a bit higher, this replicated the pattern reported in Soto and John (2017b; e.g., the depression facet of negative emotionality loaded negatively on extraversion). At the facet level, we used confirmatory factor analysis to demonstrate that model fit was good when items from each Big Five domain were considered separately.

We also obtained satisfactory levels of Cronbach's alpha for the BFI–2 domain and facet scale scores, the only exception being the responsibility facet scale. In this case, Cronbach's alpha was suppressed by Item 28, which in the exploratory principal component analysis did not show a secondary loading on agreeableness like the other three items (see Table 1). This is partly at odds with Soto and John (2017b), who reported that only one item ("Is reliable, can always be counted on") had a secondary loading on agreeableness. For the applied user in the Dutch context, we recommend caution in interpreting this specific facet until more research is available on its nature.

One notable finding was the relatively high average intercorrelation between the domain scales, which was .31 in our study, whereas Soto and John (2017b) reported only .20 in the Internet volunteer sample, and .24 in the student sample. We suspect that our higher intercorrelation is linked to our use of paid representative samples, which tend to have higher discriminant correlations than either student samples or self-selected Internet samples. To investigate this hypothesis, we analyzed additional data from a volunteer sample collected by students (with a high percentage of respondents under 30, M age =26). In this sample, the average intercorrelation was .20. Part of the reason is likely that student and volunteer samples are typically highly educated, which relates positively with factor differentiation (Rammstedt, Goldberg, & Borg, 2010). Consistent with this, in our sample the average domain correlation was higher in people with less ($r = .36$) versus more ($r = .26$) education.

Finally, our results indicated strong convergent validity of the BFI–2 domain scales scores with the IPIP Big Five scales. This suggests that the added inclusion of facets did not shift the scale scores away from the Big Five structure as represented by the IPIP items.

Study 2

Study 1 produced a Dutch version of the BFI–2 with sound psychometric properties, including a well-fitting dimensional structure, internal consistencies that were good for the domain scales and adequate-to-good for most facet scales, and high levels of convergent validity vis-à-vis an established Big Five instrument. It thus appears that our goal of translating and adapting the English BFI–2 for use in the Dutch context was achieved.

In Study 2, we set out to address the second goal of the article: Establish the predictive validity of the BFI–2 domain and facet scales. We used three approaches to do this. First, like in the original Dutch BFI paper (Denissen et al., 2008), age and gender correlates of the BFI–2 were computed and compared with established meta-analytic patterns. Second, the panel that participated in Study 1 provided a broad range of criterion variables that were also included in the Ozer and Benet-Martínez (2006) review. We used an expert panel to operationalize our hypotheses regarding the predictive validity of the domain scores. Third, a unique feature of our study was that we preregistered hypotheses regarding the expected unique predictive validity of each assessed personality facet (see Open Science Framework [OSF] at https://osf.io/gkh8j/). Resulting evidence is therefore of major importance for the bandwidth-fidelity discussion.

Method

Procedure

To compute the age and gender correlates of the BFI–2 domain and facet scale scores, we relied on the information collected at the time of the BFI–2 assessment. Other criteria were drawn from different modules from the same large-scale representative LISS panel study. This ongoing panel invites participants to fill out a different questionnaire module every month. It has modules about health, religion and ethnicity, social integration and leisure, family and household, work and schooling, personality, politics and values, financial assets, income, and housing. The files containing these criteria were merged with the responses from the BFI–2, using the data collection that was closest to the assessment of the BFI–2 (for the exact time difference, see Tables 5 and 6).

Sample

The sample was largely overlapping with the one of Study 1. However, because we merged with criterion data sets that were collected at different time points before and after the BFI–2 assessment, and because some criteria were conditional on the life situation of the participants (e.g., satisfaction with work can only be measured in participants who have a job) the actual sample size differed for each analysis (see Tables 5 and 6).

Materials

A research intern screened the coding books of all these modules, and extracted all criteria with relevance to the outcome variables reviewed in Ozer and Benet-Martínez (2006). The resulting list was screened by the intern and the first author and variables were excluded if they were (a) not included in all (yearly) waves, (b) could hardly be influenced by participants (e.g., nationality), (c) were only relevant for specific age periods (e.g., retirement), (d) were ambiguous with regard to the Ozer and Benet-Martínez criteria (e.g., "subjective standard of living" was deemed too distant from Ozer and Benet-Martínez's criterion of life satisfaction), or (e) could be seen as the lower level manifestation of a higher order construct, in which case the higher order construct was selected (e.g., "satisfaction with education" was treated as a lower level manifestation of life satisfaction). The list of criteria was discussed during three iterative rounds of iterations with the first author, until a list of 95 possible criteria was settled on.

The list with 95 criteria was then processed by the first, second, and third authors, who used it to independently predict for each facet how it would be associated with the criterion, using the response options −2 (*clear negative association*), −1 (*possible negative association*), 0 (*no association*), +1 (*possible positive association*), and +2 (*clear positive association*). Correlations between these three judges (across a vector of 95 criteria ×15 facets =1,425 entries) were relatively high, ranging between .56 and .66, $p < .001$. It therefore was justified to create an aggregate score representing the predicted ability of each facet to predict the 95 criteria, with a Cronbach's alpha of .83. When the absolute predictions were inspected for each criterion, the trait with the largest score had an average predicted score of |1.05| across criteria. In other words, as intended the raters regarded the criteria as possibly related to at least one of the Big Five domains.

We used the aggregated ratings to identify criteria against which we could assess the predictive validity of the domain scales. Specifically, we flagged criteria related to each Big Five domain scale when the absolute average prediction was at least 1, and when this prediction did not differ more than one rating point between the facets. This resulted in 44 criterion variables, which are listed in Table 5. As can be seen, many criteria were related to affective states (e.g., feeling ashamed; feeling strong), self-endorsed values (e.g., importance of politeness, importance of being open-minded), or satisfaction with life outcomes (e.g., satisfaction with life, satisfaction with leisure time).

To establish the discriminant validity of the BFI–2 facets, we derived predictions for facets that (a) were at least rated with an absolute score of 1 (corresponding to a possible positive/negative association), and (b) where the difference between the focal facet and the average prediction for the nonfocal facets of the trait was greater than one. For example, if the average prediction for the focal facet was 1.5, the average prediction for the nonfocal facets could not exceed 0.5 to be included as a hypothesis. The resulting 28 predictions are shown in Table 6. For one criterion, substance use, the LISS panel contained three indicators relating to smoking, alcohol, and drug use. Because these indicators were relatively uncorrelated, we tested predictive associations for each of them separately. Of the 28 differential predictions, 27 were preregistered at the OSF (the other one was missed due to a clerical error). Table 5 lists the descriptive statistics for each criterion. As can be seen, some variables were relatively normally distributed but others were skewed or dichotomous. We adjusted our statistical procedure to fit each distribution (Pearson product–moment correlation, Spearman rank-order correlation, or logistic regression, respectively).

Results

Demographic correlates

Age and gender correlates of the Big Five domain and facet scales are presented in Table 4. To obtain these correlates, a series of multiple regression analyses was carried out, regressing each standardized domain or facet scale score in a single regression on age, age squared, and gender. To avoid inflation of the number of tests, interactions between gender and age (squared) were not computed. Because we did not establish any formal procedure to preregister our predictions, we corrected for chance capitalization using the "BY" procedure (Benjamini & Yekutieli, 2001) in the R statistical package (R Core Team, 2017). Similarly, we computed the 99.9% confidence intervals to (approximately) match the .001 alpha level.

Table 4. Regression weights and confidence intervals for age (range =18–83), age squared, and gender predicting Big Five Inventory–2 (BFI–2) domain and facet scales.

	Age		Age squared		Female gender	
Factor/facet	b	99.9% CI	b	99.9% CI	b	99.9% CI
Extraversion	.06	[−.06, .17]	−.05	[−.16, .06]	−.03	[−.25, .19]
Sociability	.05	[−.06, .16]	−.07	[−.18, .05]	.08	[−.13, .30]
Assertiveness	.07	[−.04, .18]	−.02	[−.13, .09]	−.34*	[−.56, −.13]
Energy level	.02	[−.09, .13]	−.03	[−.14, .08]	.16	[−.05, .38]
Agreeableness	.15*	[.04, .26]	−.07	[−.18, .04]	.37*	[.16, .59]
Compassion	.14*	[.03, .24]	−.07	[−.18, .04]	.50*	[.29, .71]
Respectfulness	.13*	[.02, .24]	−.06	[−.17, .05]	.28*	[.07, .50]
Trust	.11*	[−.00, .22]	−.05	[−.16, .06]	.17	[−.05, .39]
Conscientiousness	.17*	[.06, .28]	−.12*	[−.23, −.01]	.19	[−.02, .40]
Organization	.11*	[.00, .23]	−.07	[−.18, .04]	.20*	[−.01, .42]
Productiveness	.22*	[.11, .33]	−.12*	[−.23, −.01]	.07	[−.15, .28]
Responsibility	.09	[−.02, .20]	−.12*	[−.23, −.01]	.22*	[.01, .44]
Negative emotionality	−.16*	[−.27, −.05]	.01	[−.10, .12]	.33*	[.11, .54]
Anxiety	−.15*	[−.26, −.04]	−.01	[−.12, .10]	.43*	[.22, .64]
Depression	−.13*	[−.24, −.02]	.04	[−.07, .15]	.15	[−.06, .37]
Emotional volatility	−.15*	[−.26, −.04]	−.01	[−.12, .11]	.26*	[.05, .48]
Open-mindedness	.06	[−.05, .18]	.02	[−.10, .13]	−.03	[−.25, .19]
Intellectual curiosity	−.07	[−.18, .04]	−.03	[−.15, .08]	−.17	[−.39, .05]
Aesthetic sensitivity	.16*	[.05, .27]	.07	[−.04, .18]	.12	[−.10, .34]
Creative imagination	.02	[−.05, .09]	−.03	[.06, −.11]	−.08	[.17, −.32]

Note. Age, age squared, and all BFI–2 scales were standardized before running the regressions, so corresponding b coefficients can be compared with standardized regression weights. The b coefficients indicating the association between gender and the BFI–2 scales can be compared with the Cohen's *d* effect size difference between men and women. Positive correlations mean that women scored higher than men.
*$p < .001$.

As can be seen in Table 4, established correlations between the Big Five traits and age were broadly replicated using the domain scales: We found age-related increases in agreeableness and conscientiousness, and decreases in negative emotionality. For conscientiousness, a negative quadratic trend was found: The increase in conscientiousness decelerated somewhat with age and even decreased slightly after age 60. Finally, we replicated established gender differences on the level of domain scales, with levels of agreeableness and negative emotionality being higher for women. As can be seen in Table 4, the effect size of these differences was small to moderate.

Table 4 shows that the significant age and gender correlates of the facets largely mirrored those of the domain scales, with some interesting exceptions (for a graphical depiction of the age trends, see Figure S.1 of the supplementary materials). To begin, there was no significantly positive age correlation with the responsibility facet of conscientiousness, running counter to the maturation pattern found for the other conscientiousness facets. A significant quadratic age effect was absent for the conscientiousness facet of organization, whereas it was observed for the other facets. Furthermore, even though generally no age trend was found for open-mindedness, for the facet of aesthetic sensitivity an age-related increase was found. Regarding gender differences, some facets demonstrated significance, whereas the pattern obtained with the domain scales did not. This was especially true for the assertiveness facet of extraversion, which was lower for women. In contrast, the organization and responsibility facets of conscientiousness were higher for women. For agreeableness and negative emotionality, the generally higher scores for women were not found for the trust and depression facets, respectively.

Predictive validity of domain scales

Table 5 lists all predictive associations for the BFI–2 domain scales. The raw rating data and the script through which we arrived at our predictions is shown at the OSF page (see https://osf.io/nwtx7/). We did not originally upload these preregistrations because the initial focus was on the facet predictions. Because we had nevertheless specified our predictions in advance, we applied the established criterion for statistical significance of $p < .05$. As can be seen, of all 44 predictions, the absolute association of the corresponding domain scale with the criterion was indeed stronger than the maximum absolute association across the other four domain scales in 38 cases (86%). In one case, membership in a sports club was not associated with extraversion (as predicted) but instead with agreeableness (positively) and negative emotionality (negatively). In the other case, the self-rated importance of being open-minded was indeed, as predicted, associated with open-mindedness—but, unexpectedly, more strongly with extraversion and agreeableness. Overall, the predicted correlations were mostly small or moderate. The average (absolute) predictive association was .31 across all predicted criteria (displayed in bold in Table 5). By comparison, the maximum absolute predictive association across all nonpredicted domain scales was .23. In sum, results therefore supported the predictive validity of the BFI–2 domain scales.

Discriminant validity of facet scales

Table 6 lists all predictive associations of the BFI–2 facets. The conventional level of .05 for statistical significance was applied for these preregistered hypotheses. As can be seen, the predicted facet was significantly correlated with the validity criterion in 21 of 28 cases. We did not count substance use as a significant result because it was significantly

Table 5. Predictive validity of the Big Five Inventory–2 (BFI–2) domain scales and background statistics for the criteria studied.

								Ext.		Agr.		Con.		Neg.		Ope.		
Outcome variable per domain	Time	*n*	Range	*M*	*SD*	Skew	Kurtosis	Coef	CI	Coef	CI	Coef	CI	Coef	CI	Coef	CI	Hit
Predictions for extraversion	−3	739	1 – 7	4.52	1.4	−0.53	0.06	**.29***	[.2, .39]	.14*	[.03, .23]	.20*	[.08, .29]	−.24*	[−.33, −.15]	.10*	[−.01, .22]	1
Feeling enthusiastic[P]																		
Feeling good[P]	−3	743	1 – 7	5.66	0.93	−1.16	2.2	**.26***	[.15, .35]	.18*	[.07, .29]	.19*	[.09, .31]	−.40*	[−.49, −.31]	.10*	[−.01, .21]	1
Feeling lonely[S]	−10	642	0.5 – 2.5	0.82	0.41	1.53	1.99	**−.25***	[−.36, −.13]	−.16*	[−.26, −.04]	−.13*	[−.24, −.03]	.27*	[.15, .37]	−.01	[−.12, .11]	1
Feeling strong[P]	−3	739	1 – 7	4.76	1.38	−0.69	0.32	**.30***	[.19, .4]	.16*	[.06, .26]	.23*	[.13, .33]	−.35*	[−.45, −.24]	.11*	[.01, .21]	1
Importance of happiness[S]	−3	739	1 – 7	6.14	1.02	−1.58	3.48	**.22***	[.13, .33]	.18*	[.07, .27]	.14*	[.04, .24]	−.10*	[−.2, .01]	.10*	[−.02, .2]	1
Membership in a sports club[L]	−10	645	0 – 1	0.35	0.48	0.63	−1.6	**.19**	[−.05,.44]	.39*	[.08,.69]	.09	[−.17,.34]	−.38*	[−.63, −.14]	−.29	[−.53, −.05]	0
Satisfaction with leisure time[P]	−10	632	0 – 10	7.26	1.59	−0.83	1.63	**.21***	[.08, .32]	.11*	[−.02, .24]	.13*	[.01, .26]	−.27*	[−.38, −.15]	.05	[−.07, .16]	1
Predictions for agreeableness	−3	727	0 – 10	6	2.18	−0.88	0.39	.06	[−.05, .17]	**.23***	[.11, .33]	.01	[−.1, .14]	−.20*	[−.32, −.09]	.19*	[.09, .3]	1
Belief that people can be trusted[P]																		
Feeling hostile[S]	3	739	1 – 7	1.69	1.15	1.99	3.88	−.12*	[−.21, −.01]	**−.29***	[−.38, −.18]	−.21*	[−.31, −.09]	.28*	[.18, .37]	−.19*	[−.29, −.09]	1
Feeling irritable[S]	−3	739	1 – 7	2.55	1.61	0.84	−0.33	−.15*	[−.26, −.04]	**−.28***	[−.38, −.18]	−.23*	[−.34, −.14]	.42*	[.32, .5]	−.17*	[−.26, −.06]	1
Importance of forgivingness[P]	−3	739	1 – 7	5.56	1.19	−0.83	0.69	.13*	[.01, .23]	**.38***	[.26, .48]	.23*	[.1, .34]	−.14*	[−.24, −.02]	.14*	[.05, .26]	1
Importance of politeness[S]	−3	739	1 – 7	5.9	1.08	−1.16	1.76	.21*	[.11, .29]	**.27***	[.15, .37]	.26*	[.16, .38]	−.10*	[−.21, 0]	.06	[−.04, .18]	1
Importance of sincereness and truthfulness[S]	−3	739	1 – 7	6.51	0.86	−3.03	13.13	.15*	[.04, .25]	**.35***	[.25, .43]	.21*	[.12, .31]	−.14*	[−.24, −.03]	.14*	[.04, .23]	1
Sense of connectedness to other people[P]	−3	739	1 – 7	4.63	1.56	−0.47	−0.4	.25*	[.15, .35]	**.29***	[.18, .4]	.18*	[.08, .28]	−.19*	[−.3, −.07]	.05	[−.04, .15]	1
Predictions for conscientiousness	−3	739	1 – 7	5.32	1.25	−0.62	0.13	.17*	[.06, .29]	.10*	[−.02, .21]	**.35***	[.23, .46]	−.13*	[−.24, −.02]	−.01	[−.12, .1]	1
Importance of being self-controlled[P]																		
Importance of responsibility[S]	−3	739	1 – 7	6.24	0.92	−1.89	5.98	.21*	[.12, .32]	.27*	[.17, .37]	**.27***	[.16, .37]	−.18*	[−.29, −.08]	.16*	[.06, .26]	0
Importance of working hard[P]	−3	739	1 – 7	5.64	1.17	−0.95	1.15	.23*	[.14, .33]	.18*	[.07, .31]	**.32***	[.21, .44]	−.21*	[−.31, −.08]	.01	[−.1, .12]	1
Feeling determined[P]	−3	739	1 – 7	4.49	1.58	−0.53	−0.26	.24*	[.14, .34]	.11*	[0, .21]	**.27***	[.17, .36]	−.24*	[−.33, −.13]	.10*	[−.01, .22]	1
Importance of a sense of accomplishment[P]	−3	739	1 – 7	4.87	1.31	−0.49	0.11	.25*	[.13, .37]	.02	[−.12, .15]	**.14***	[.02, .25]	−.17*	[−.28, −.05]	.05	[−.07, .17]	0
Predictions for negative emotionality																		0
Expecting negative changes to the total net income of household[P]	−2	629	1 – 5	2.94	0.7	−0.04	1.27	−.08*	[−.19, .04]	−.04	[−.17, .09]	.01	[−.13, .14]	**.06**	[−.07, .18]	0	[−.16, .12]	1
Feeling ashamed[S]	−3	739	1 – 7	1.92	1.3	1.6	2.15	−.18*	[−.29, −.08]	−.24*	[−.34, −.13]	−.19*	[−.3, −.09]	**.32***	[.22, .43]	−.21*	[−.29, −.11]	1
Feeling distressed[S]	−3	739	1 – 7	2.96	1.7	0.57	−0.69	−.11*	[−.23, 0]	−.17*	[−.29, −.06]	−.11*	[−.23, .02]	**.34***	[.25, .44]	−.07	[−.18, .03]	1
Feeling good[P]	−3	743	1 – 7	5.66	0.93	−1.16	2.2	.26*	[.15, .36]	.18*	[.05, .29]	.19*	[.08, .30]	**−.40***	[−.5, −.30]	.10*	[−.01, .20]	1
Feeling guilty[S]	−3	739	1 – 7	1.89	1.3	1.66	2.31	−.18*	[−.29, −.08]	−.25*	[−.35, −.16]	−.23*	[−.33, −.11]	**.32***	[.23, .42]	−.18*	[−.28, −.08]	1
Feeling hostile[S]	−3	739	1 – 7	1.69	1.15	1.99	3.88	−.12*	[−.22, −.01]	−.29*	[−.37, −.18]	−.21*	[−.3, −.11]	**.28***	[.16, .39]	−.19*	[−.28, −.09]	1
Feeling irritable[S]	−3	739	1 – 7	2.55	1.61	0.84	−0.33	−.15*	[−.24, −.04]	−.28*	[−.39, −.18]	−.23*	[−.33, −.14]	**.42***	[.32, .51]	−.17*	[−.27, −.06]	1
Feeling jittery[S]	−3	739	1 – 7	2.24	1.52	1.22	0.64	−.13*	[−.23, −.03]	−.25*	[−.36, −.16]	−.22*	[−.32, −.11]	**.40***	[.3, .49]	−.17*	[−.27, −.05]	1
Feeling lonely[S]	−3	642	0.5 – 2.5	0.82	0.41	1.53	1.99	−.25*	[−.34, −.11]	−.16*	[−.27, −.04]	−.13*	[−.23, −.01]	**.27***	[.16, .37]	−.01	[−.11, .09]	1
Feeling nervous[S]	−3	739	1 – 7	2.2	1.49	1.22	0.63	−.14*	[−.24, −.04]	−.24*	[−.34, −.14]	−.21*	[−.32, −.11]	**.40***	[.3, .51]	−.14*	[−.25, −.03]	1
Feeling scared[S]	−3	739	1 – 7	1.81	1.28	1.81	2.83	−.15*	[−.25, −.05]	−.27*	[−.37, −.17]	−.23*	[−.31, −.11]	**.38***	[.29, .47]	−.20*	[−.3, −.09]	1
Feeling strong[P]	−3	739	1 – 7	4.76	1.38	−0.69	0.32	.30*	[.18, .39]	.16*	[.03, .26]	.23*	[.14, .33]	**−.35***	[−.45, −.25]	.11*	[0, .21]	1
Feeling upset[S]	−3	739	1 – 7	2	1.33	1.46	1.61	−.13*	[−.23, −.02]	−.29*	[−.38, −.19]	−.21*	[−.31, −.1]	**.38***	[.28, .47]	−.21*	[−.3, −.12]	1
General happiness[P]	−3	722	0 – 10	7.43	1.29	−1.33	3.72	.23*	[.13, .35]	.15*	[0, .28]	.19*	[.06, .31]	**−.34***	[−.45, −.24]	.06	[−.03, .14]	1
Life satisfaction[P]	−3	743	1 – 7	5.02	1.11	−0.83	0.58	.24*	[.13, .35]	.13*	[−.01, .24]	.18*	[.08, .29]	**−.38***	[−.46, −.29]	.04	[−.06, .15]	1
Satisfaction with leisure time[P]	−10	632	0 – 10	7.26	1.59	−0.83	1.63	.21*	[.09, .32]	.11*	[−.01, .21]	.13*	[.02, .25]	**−.27***	[−.39, −.15]	.05	[−.06, .16]	1
Satisfaction with work[S]	−3	297	0 – 10	7.38	1.3	−1.06	2.51	.21*	[.05, .39]	.22*	[.05, .38]	.21*	[.05, .36]	**−.31***	[−.46, −.15]	−.04	[−.22, .12]	1
Self-esteem[P]	−3	742	1.1 – 7	5.47	0.97	−0.69	0.21	.40*	[.31, .5]	.28*	[.16, .38]	.29*	[.19, .39]	**−.58***	[−.65, −.51]	.16*	[.06, .27]	1
Subjective rating of being healthy[P]	3	740	1 – 5	3.14	0.81	0.18	0.28	.12*	[.03, .21]	.05	[−.07, .18]	.06	[−.06, .19]	**−.24***	[−.34, −.13]	−.01	[−.11, .1]	1

(*continued*)

Table 5. Continued.

Outcome variable per domain	Time	*n*	Range	*M*	*SD*	Skew	Kurtosis	Ext.		Agr.		Con.		Neg.		Ope.		Hit
								Coef	CI	Coef	CI	Coef	CI	Coef	CI	Coef	CI	
Predictions for open-mindedness	3	554	0 – 10	5.2	2.16	−0.22	−0.43	.03	[−.08, .14]	−.14*	[−.25, −.01]	.04	[−.08, .16]	−.09*	[−.2, .03]	**−.23***	[−.35, −.1]	1
Being a conservative voter[P]																		
Enjoying intellectual activity and thinking about problems[P]	−3	739	1 – 7	4.15	1.67	−0.31	−0.67	.10*	[−.01, .21]	.01	[−.09, .13]	.05	[−.04, .15]	−.15*	[−.27, −.04]	**.35***	[.26, .45]	0
Feeling inspired[P]	−3	739	1 – 7	3.97	1.59	−0.26	−0.61	.28*	[.19, .38]	.12*	[.02, .22]	.17*	[.07, .27]	−.19*	[−.27, −.09]	**.24***	[.14, .33]	0
Importance of being open-minded[P]	−3	739	1 – 7	5.66	1.15	−0.91	0.87	.32*	[.22, .43]	.29*	[.17, .4]	.17*	[.05, .28]	−.13*	[−.25, −.01]	**.15***	[.06, .26]	0
Importance of creativity[P]	−3	739	1 – 7	4.95	1.39	−0.42	−0.24	.26*	[.15, .37]	.14*	[.05, .25]	.15*	[.03, .25]	−.11*	[−.23, .01]	**.44***	[.35, .53]	1
Importance of intellectuality[P]	−3	739	1 – 7	4.91	1.31	−0.47	0.05	.18*	[.07, .31]	.04	[−.08, .16]	.07	[−.05, .18]	−.12*	[−.22, 0]	**.27***	[.17, .37]	1

Note. Ext. = extraversion; Agr. = agreeableness; Con. = conscientiousness; Neg. = negative emotionality; Ope. = open-mindedness. Superscript P = Pearson correlation; S = Spearman rank-order correlation; L = logistic regression coefficient. Statistics that are indicative of the hypotheses are displayed in bold. Time refers to the time (in months) of assessment of the criterion variable, relative to the BFI–2, with negative numbers referring to an earlier time of assessment (antediction) and positive numbers to a later time point (prediction). *n* refers to the sample size for the bivariate association, which differed based on the module the criterion was included in, as well as other factors (e.g., satisfaction with work was only assessed for participants with paid employment). Hit refers to the comparison of predicted and actual association, with a value of 0 when at least one (absolute) nonpredicted Big Five association was stronger than the absolute predicted association, and 1 if otherwise. In case multiple predictors were hypothesized to be associated with a criterion and one of the predictors indeed produced the strongest association, the "hit" variable was still coded as 1.

*$p < .05$.

associated with responsibility in only one of the three cases (i.e., for drug use). Moreover, in 19 out of 28 cases the absolute correlations between the predicted facet with the validity criterion were higher than with the other two same-domain facets. Overall, the pattern of hits versus misses was most favorable for the facets of agreeableness and openness, whereas the picture was more mixed for the other traits.

As registered at OSF, we tallied the cases where the predicted facet was indeed the strongest predictor in the expected direction (a "hit") versus cases in which this was not the case (a "miss"). We counted substance use as a hit, because it was most strongly associated with responsibility in two of the three cases (i.e., for smoking and drug use). The resulting Fisher's exact test indicated that the pattern of hits (19 out of 28 cases; 67.9%) was higher than the number of hits that would be predicted by chance (33.3%), with an odds ratio of 10.61, $p < .001$. As can be seen in Table 6, however, in no case was the 95% confidence interval nonoverlapping between the facets, so our positive finding concerns the overall pattern of facet–criteria associations, rather than specific bivariate contrasts.

Discussion

The results of Study 2 indicated that the Big Five domain scales indeed were associated with our predicted external correlates, being robustly correlated with age and gender differences and a range of external criteria. The correlations of the facets with demographic criteria were broadly similar to those of the domain scales, but with some interesting exceptions (e.g., the age-related increase in aesthetic sensitivity that was not found for the overall domain of open-mindedness). Finally, we showed that the majority of criteria were best predicted by preregistered facets, supporting their discriminant validity.

General discussion

The Dutch BFI–2

This study pursued two goals. We examined whether an adaptation of the English BFI–2 for use in the Dutch language was successful in terms of reliability, structural validity, and convergent validity vis-à-vis an established Big Five instrument. This first goal was achieved: After a series of iterative improvements (described in the supplementary materials), results of Study 1 indicated that our final Dutch BFI–2 version replicated the good psychometric properties of the English original, and shows the validity and usability of the instrument beyond the original language. In addition, these results are of relevance to the broader international personality literature in a number of ways.

Predictive validity of domains and facets

The second main goal of this research was to test the predictive validity of the domain and facet scales of the BFI–2. Study 2 addressed this goal by examining associations of the

Table 6. Predictive validity of the Big Five Inventory–2 (BFI–2) facet scales and background statistics for the criteria studied.

									Facet 1		Facet 2		Facet 3		
Outcome variable	Time	*n*	Range	*M*	*SD*	Skew	Kurtosis	Prediction	Coef	CI	Coef	CI	Coef	CI	Hit
Prediction for extraversion facets									Sociability		Assertiveness		Energy level		
Importance of true friendship[S]	−3	739	1 – 7	5.96	1.14	−1.32	2.05	E1 (sociability)	**.21***	[.14, .28]	.16*	[.09, .24]	.21*	[.14, .28]	0
Sense of connectedness to other people[P]	−3	739	1 – 7	4.63	1.56	−0.47	−0.4	E1 (sociability)	**.23***	[.16, .29]	.14*	[.07, .22]	.26*	[.20, .33]	0
Number of close relationships[S]	−10	586	0 – 5	2.58	1.78	−0.12	−1.24	E1 (sociability)	**.07**	[−.01, .16]	.04	[−.04, .12]	.10*	[.02, .18]	0
Feeling determined[P]	−3	739	1 – 7	4.49	1.58	−0.53	−0.26	E2 (assertiveness)	.16*	[.09, .23]	**.24***	[.17, .31]	.21*	[.11, .28]	1
Importance of self-respect[S]	−3	739	1 – 7	6.14	1	−1.49	3.27	E2 (assertiveness)	.18*	[.11, .25]	**.18***	[.11, .24]	.24*	[.18, .31]	0
Time spent with physical activities[P]	3	739	0 – 7	2.83	1.6	0.18	−0.51	E3 (energy level)	.06	[−.02, .13]	.02	[−.05, .08]	**.10***	[.04, .17]	1
Feeling excited[P]	−3	739	1 – 7	2.86	1.55	0.47	−0.65	E3 (energy level)	.07	[−.01, .13]	.02	[−.05, .10]	**.07**	[−.00, .13]	0
Importance of an exciting life[P]	−3	739	1 – 7	4.46	1.45	−0.31	−0.26	E3 (energy level)	.26*	[.19, .33]	.23*	[.15, .30]	**.28***	[.20, .35]	1
Prediction for agreeableness facets									Compassion		Respectfulness		Trust		
Importance of being loving[S]	−3	739	1 – 7	6.08	1.03	−1.45	3.02	A1 (compassion)	**.42***	[.35, .48]	.32*	[.25, .39]	.24*	[.19, .31]	1
Importance of being helpful[S]	−3	739	1 – 7	6	0.99	−1.24	2.49	A1 (compassion)	**.41***	[.35, .47]	.30*	[.22, .36]	.25*	[.17, .32]	1
Providing care to someone else[L]	−10	645	0 – 1	0.21	0.41	1.4	−0.03	A1 (compassion)	**.56***	[.24,.88]	.37	[.05,.70]	.26	[−.04,.56]	1
Having quarrels with partner[S]	−23	461	1 – 3	1.4	0.41	1.06	0.84	A2 (respectfulness)	−.04	[−.13, .06]	**−.12***	[−.21, −.01]	−.04	[−.12, .05]	1
Prediction for conscientiousness facets									Organization		Productiveness		Responsibility		
Importance of cleanness[P]	−3	739	1 – 7	5.24	1.31	−0.56	−0.12	C1 (organization)	**.42***	[.36, .48]	.36*	[.30, .42]	.29*	[.22, .35]	1
Taxable income[P]	−2	463	2 – 10	6.41	2.03	−0.24	−0.53	C2 (productiveness)	−.03	[−.12, .07]	**−.01**	[−.11, .08]	.01	[−.09, .11]	0
Importance of capability[P]	−3	739	1 – 7	5.45	1.12	−0.64	0.45	C2 (productiveness)	.10*	[.01, .17]	**.24***	[.16, .31]	.12*	[.04, .18]	1
Having a high job qualification[S]	−4	538	1 – 27	15.47	6.63	−0.19	−0.97	C2 (productiveness)	−.02	[−.09, .04]	**−.05**	[−.14, .03]	.01	[−.06, .08]	0
Substance use: Tobacco[L]	3	740	0 – 1	0.54	0.5	−0.17	−1.97	C3 (responsibility)	−.06	[−.23,.10]	.13	[−.06,.32]	**−.23**	[−.46,.01]	1
Substance use: Alcohol[P]	3	740	0 – 7	3.46	2.2	−0.05	−0.93	C3 (responsibility)	−.04	[−.11, .03]	.08*	[.02, .16]	**.00**	[−.08, .08]	0
Substance use: Drugs[L]	3	740	0 – 1	0.05	0.21	4.37	17.12	C3 (responsibility)	−1.06*	[−1.47,−.64]	−1.10*	[−1.57,−.62]	**−1.71***	[−2.34,−1.07]	1
Being religious[L]	−12	638	0 – 1	0.35	0.48	0.65	−1.58	C3 (responsibility)	.07	[−.12,.26]	.40*	[.17,.63]	**.42***	[.14,.69]	1
Providing care to someone else[L]	−10	645	0 – 1	0.21	0.41	1.4	−0.03	C3 (responsibility)	.02	[−.20,.24]	.25	[−.01,.52]	**.24**	[−.08,.56]	0
Prediction for negative emotionality facets									Anxiety		Depression		Emotional volatility		
Expecting positive changes to the total net income[P]	−2	629	1 – 5	3.06	0.7	0.04	1.27	N1 (anxiety)	**−.09***	[−.18, −.02]	−.08*	[−.16, .00]	.02	[−.06, .10]	1
Feeling afraid[S]	−3	739	1 – 7	1.89	1.34	1.72	2.49	N1 (anxiety)	**.34***	[.28, .40]	.33*	[.28, .40]	.30*	[.23, .36]	1
Feeling excited[P]	−3	739	1 – 7	2.86	1.55	0.47	−0.65	N2 (depression)	.07	[−.01, .14]	**.05**	[−.01, .12]	.08*	[.01, .16]	0
Prediction for open-mindedness facets									Intellectual curiosity		Aesthetic sensitivity		Creative imagination		
Feeling interested[P]	−3	739	1 – 7	5.22	1.32	−0.95	0.97	O1 (intellectual cur.]	**.20***	[.13, .27]	.12*	[.04, .18]	.18*	[.11, .25]	1
Importance of logic[P]	−3	739	1 – 7	5.25	1.16	−0.52	0.21	O1 (intellectual cur.]	**.17***	[.11, .24]	.08*	[.01, .15]	.15*	[.07, .23]	1
Interest in following the news[P]	−8	611	−1.76 – 1.52	0.05	0.77	0.05	−0.46	O1 (intellectual cur.]	**.36***	[.29, .43]	.28*	[.20, .36]	.21*	[.12, .29]	1
Having a high job qualification[S]	−4	538	1 – 27	15.47	6.63	−0.19	−0.97	O1 (intellectual cur.]	**.30***	[.24, .37]	.16*	[.09, .23]	.12*	[.05, .21]	1
Importance of a world at peace[S]	−3	739	1 – 7	5.87	1.34	−1.3	1.38	O2 (aesthetic sens.]	.08*	[.00, .15]	**.16***	[.09, .23]	.08*	[.02, .16]	1
Importance of a world of beauty[P]	−3	739	1 – 7	4.83	1.34	−0.43	−0.01	O2 (aesthetic sens.]	.07*	[−.01, .14]	**.17***	[.11, .25]	.11*	[.04, .18]	1

Note. Superscripts after each outcome variable refers to the statistic that was used to test the prediction, with P referring to Pearson correlations, S to Spearman rank-order correlations, and L to logistic regression coefficients. Statistics that are indicative of the hypotheses are displayed in bold. Time refers to the time (in months) of assessment of the criterion variable, relative to the BFI–2, with negative numbers referring to an earlier time of assessment (antediction) and positive numbers to a later time point (prediction). *n* refers to the sample size for the bivariate association, which differed based on the module the criterion was included in, as well as other factors (e.g., satisfaction with work was only assessed for participants with a paid employment). Hit refers to the comparison of predicted and actual association, with a value of 0 when nonpredicted Big Five associations were stronger, and a value of 1 when the predicted association was strongest. Predictions for importance of helpfulness were not preregistered due to a clerical error but evidence can be obtained from the first author that this criterion should also have been included in the list of predictions.

*$p < .05$.

domain and facet scales with age, gender, and a variety of external criteria. With the domain scales, we replicated established age and gender correlates of the Big Five, such as age-related increases in agreeableness, conscientiousness and emotional stability (i.e., decreases in negative emotionality), and higher levels of agreeableness and negative emotionality for women.

The facets generally displayed similar age and gender correlates when compared to the overarching domain scales, but with some interesting exceptions. For example, aesthetic sensitivity increased with age, whereas the overarching open-mindedness domain scale did not. This pattern runs counter to the published age-related decreases in cultural activity by Schwaba, Luhmann, Denissen, Chung, and Bleidorn (2018), based on the LISS sample. Possibly, the latter decrease represents an age-related decline in mobility that limits cultural participation, disguising an increase with age of the desire for aesthetic stimulation. Interestingly, no difference between extraversion facets was found in terms of age correlates, which runs counter to the Roberts et al. (2006) finding that social dominance increased with age, which might be reflected in increasing assertiveness levels.

Finally, the facets also uncovered interesting gender differences in personality traits that were not evident at the domain level. For example, overall extraversion levels did not differ between men and women, but men did report higher levels of assertiveness. This latter result replicated a finding by Weisberg et al. (2011) that men scored higher on the extraversion facet of dominance but it runs counter to the findings of Soto and John (2017b), who did not find any gender difference in this facet. More research into possible explanations of this discrepancy is needed. For negative emotionality, higher levels for women were found, except for depression. This replicated findings by Soto and John (2017b), who also reported a lack of gender differences in this facet. Perhaps the attenuated gender difference in depression is due to its secondary loading on extraversion, for which gender differences are not found across the board. More research is needed to substantiate this conclusion, however.

To the best of our knowledge, Study 2 was the first to systematically scan all variables of a large-scale multifaceted panel study and to use a priori ratings by experts to pinpoint and preregister predictive associations. One set of 44 associations was thus derived for the domain scales. In these cases, experts predicted that all BFI–2 facets within a certain Big Five domain would be associated with the criterion. In 38 of these 44 cases, the correlations between this domain scale and the criterion were highest in comparison to the predictive associations of the other four (nonpredicted) Big Five domain scales.

The average predictive association (excluding the logistic regression coefficients) was .31 across all domain scales and criteria. In other words, almost 10% of the variance in our criteria was predicted by the focal domain scale. When all traits were entered simultaneously as predictors in a multiple regression, the R^2 increased to .14, so an additional 4% of the variance was predicted by the nonpredicted domain scales. The percentage of variance thus explained was smaller than that of Soto and John (2017b), who found that 27% of the variance was explained by the BFI–2 domain scales. The fact that we assessed predictor and criterion variables at different time points might have attenuated associations because both predictor and criterion might undergo changes over time. Also, many criterion variables were assessed with single items and therefore had relatively modest reliabilities (for an overview of test–retest reliabilities, see Tables S.12 and S.13). Future work might therefore establish stronger predictive validities by focusing on concurrent associations with more reliable criteria.

In this study, we also tested the unique predictive validity of the facets. Soto and John (2017b) focused on the total amount of variance that was uniquely predicted by the facets. In their study, the facets as a whole explained an additional 6% of the variance over and above the domain scales. In our study, we took a different approach by preregistering specific hypotheses regarding facets that should be more strongly associated with 28 criteria than the other facets of the corresponding domain. In 19 of these cases, the predicted facet indeed showed a stronger trait–criterion association than did the nonpredicted facets, whereas based on chance only 9 hits would be predicted. Our approach of testing the discriminant predictive validity of the facets has some advantages over testing the incremental predictive validity of the facets as a block, over and above the predictive validity of the corresponding domains (as done by Soto & John, 2017b). In the latter case, the domain scales are linear combinations of the facets, and also the increase in number of predictors comes with the risk of chance capitalization. Because we avoided this risk, it cannot be claimed, as has been done in the past (Ones & Viswesvaran, 1996), that the unique predictive validity of the facets is merely due to chance capitalization. Rather, several facets were related to criteria in a way that is in line with their conceptualization. For example, compassion was the only agreeableness facet that predicted providing care to someone else.

That said, it was clear that the facets were not dramatically distinct from each other in predicting our handpicked criteria (i.e., those variables that experts flagged as specifically associated with one particular facet). As can be seen in Table 6, in no case were confidence intervals of the predicted association outside the confidence intervals of the nonpredicted associations.[2] It can thus be questioned whether this limited increase in predictive validity offsets the facets' attenuated reliability in assessment contexts. As became evident from our correlational analyses (see Table S.4), most correlations between facets of the same domain were in the .5 to .7 range, meaning that a small but nontrivial proportion of people will have pronouncedly divergent facet profiles (e.g., high assertiveness with low sociability and energy levels). Future research in large samples using

[2]It should be noted that the relevant comparison is between dependent correlations (i.e., having the criterion variable in common) and that some pairwise comparisons actually were statistically significant using a test of the difference between two dependent correlations with one variable in common (Lee & Preacher, 2013).

person-centered techniques could investigate whether such profiles can be reliably identified, and whether they are associated with unique predictive outcomes.

Independent of their predictive specificity, however, the identification of facets has another clear advantage: It delineates the spectrum of each Big Five domain. This way, it can be ascertained that the item content of the domain scales is representative of the broader personality space. For example, it is relatively easy to create short personality scales using highly synonymous items, but such a solution sacrifices the breadth of the resulting scale (Denissen, Geenen, Selfhout, and van Aken, 2008). In contrast, the short scales of the BFI–2 were carefully created by sampling items from each facet domain. As can be seen in Table 2, this nevertheless resulted in adequate Cronbach's alphas. For researchers who are interested in capturing the entire Big Five space when there are severe constraints on assessment time, use of the resulting short scales is therefore recommended. For researchers who want to focus on more circumscribed personality dimensions, the use of the facets (which mostly had adequate Cronbach's alphas even though measured with only four items) is recommended.

Strengths, limitations, and future research

This article has a number of key strengths. It relied on multiple samples, including a large and nationally representative data set. Furthermore, the items of the English BFI–2 version were carefully and systematically translated and adapted in an iterative procedure. Also, we used a systematic approach to derive hypotheses regarding the predictive validity of the domain and facet scales, and preregistered our hypotheses for the facets. This resulted in a broad and diverse array of validation criteria. Our results indicated replicability of the original BFI–2 by showing good psychometric properties of the Dutch BFI–2, and supported the predictive specificity of the BFI–2 facets.

That said, some limitations of our approach make it necessary to conduct further research on additional properties of the BFI–2. For example, the BFI is limited to the Big Five framework and does not contain a scale directly tapping into honesty/humility from the HEXACO framework. Future studies might therefore consider adding a three-facet honesty/humility domain to the BFI–2. Second, future research might benefit from assessing additional validation criteria. For example, it would be important to assess external criteria via peer ratings and objective tests or behavior data instead of self-report. Furthermore, although we established the convergent validity of the BFI–2 scales vis-à-vis Goldberg's IPIP scales, additional work might compare the Dutch BFI–2 scales to other instruments that also assess facets, such as the NEO PI–R (Costa & McCrae, 1992; see Soto & John, 2017b), for convergence of the English BFI–2 facets and the NEO PI–R facets). Third, our panel of experts did not identify distinctive criteria for some of the facets in the LISS data set (the trust facet of agreeableness, emotional volatility facet of negative emotionality, and creative imagination facet of open-mindedness). Future research should therefore focus on these facets, ideally preregistering and then testing hypotheses regarding their predictive validity.

Conclusion

This study set out to develop and test a Dutch adaptation of the BFI–2, and to further investigate the validity of the BFI–2 domains and facets. Overall, we replicated the support that Soto and John (2017b) reported for the structural validity of the English BFI–2, as well as the convergent validity and internal consistency of its domain and facet scales. We therefore recommend use of the BFI–2 in studies with Dutch-speaking participants, as well as adaptation to other languages and cultural contexts. Furthermore, our study supported the predictive validity of the Big Five domain scales using a broad array of criteria from a large and representative longitudinal study. Importantly, our study was the first to test preregistered hypotheses regarding the incremental validity of specific facets. Our results indicated that the preregistered facets were indeed more often the strongest predictor of our selected criterion variables than would be expected by chance, although the added advantage was often subtle. More research is needed to establish additional psychometric properties of the BFI–2 and address the role of facets in personality assessment and theory.

Open Practices

This article has earned the Center for Open Science badges for Open Data, Open Materials and Preregistered through Open Practices Disclosure. The preregistered hypotheses are openly accessible at https://osf.io/gkh8j/. The data and materials are openly accessible at https://osf.io/nwtx7. To obtain the author's disclosure form, please contact the Editor.

References

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4), 1165–1188.

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

Cronbach, L. J., & Gleser, G. C. (1957). *Psychological tests and personnel decisions*. Urbana, IL: University of Illinois Press.

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

Denissen, J. J. A., Geenen, R., Van Aken, M. A. G., Gosling, S. D., & Potter, J. (2008). Development and validation of a Dutch translation of the Big Five Inventory (BFI). *Journal of Personality Assessment*, 90(2), 152–157.

Denissen, J. J. A., & Penke, L. (2008). Motivational individual reaction norms underlying the Five-Factor model of personality: First steps towards a theory-based conceptual framework. *Journal of Research in Personality*, 42(5), 1285–1302.

De Raad, B., & Peabody, D. (2005). Cross-culturally recurrent personality factors: analyses of three factors. *European Journal of Personality*, 19(6), 451–474.

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology*, 93(5), 880–896.

Feingold, A. (1994). Gender differences in personality: A meta-analysis. *Psychological Bulletin*, 116(3), 429–456.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, 40(1), 84–96.

Hagger-Johnson, G. E., & Whiteman, M. C. (2007). Conscientiousness facets and health behaviors: A latent variable modeling approach. *Personality and Individual Differences*, 43(5), 1235–1245.

Jamshidian, M., Jalal, S., & Jansen, C. (2014). *MissMech: An R package for testing homoscedasticity, multivariate normality, and missing completely at random (MCAR).* Retrieved from http://www.jstatsoft.org/v56/i06/

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big-Five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (3rd ed., pp. 114–158). New York, NY: Guilford.

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* (Vol. *2*, pp. 102–138). New York: Guilford Press.

Klimstra, T. A., Luyckx, K., Goossens, L., Teppers, E., & De Fruyt, F. (2013). Associations of identity dimensions with big five personality domains and facets. *European Journal of Personality*, 27(3), 213–221.

Klimstra, T. A., Luyckx, K., Hale, W. W., III., & Goossens, L. (2014). Personality and externalizing behavior in the transition to young adulthood: the additive value of personality facets. *Social Psychiatry and Psychiatric Epidemiology*, 49(8), 1319–1333.

Lee, I. A., & Preacher, K. J. (2013). September). Calculation for the test of the difference between two dependent correlations with one variable in common [Computer software]. Available from http://quantpsy.org/.

Lehmann, R., Denissen, J. J. A., Allemand, M., & Penke, L. (2013). Age and gender differences in motivational manifestations of the big five from age 16 to 60. *Developmental Psychology*, 49(2), 365–383.

MacCann, C., Duckworth, A. L., & Roberts, R. D. (2009). Empirical identification of the major facets of conscientiousness. *Learning and Individual Differences*, 19(4), 451–458.

McCrae, R. R., & Terracciano, A. & Personality Profiles of Cultures Project (2005). Universal features of personality traits from the observer's perspective: data from 50 cultures. *Journal of Personality and Social Psychology*, 88(3), 547–561.

Mund, M., & Neyer, F. J. (2014). Treating personality-relationship transactions with respect: narrow facets, advanced models, and extended time frames. *Journal of Personality and Social Psychology*, 107(2), 352–368.

Ones, D. S., & Viswesvaran, C. (1996). Bandwidth-fidelity dilemma in personality measurement for personnel selection. *Journal of Organizational Behavior*, 17(6), 609–626.

Ozer, D. J., & Benet-Martínez, V. (2006). Personality and the prediction of consequential outcomes. *Annual Review of Psychology*, 57, 401–421.

Paunonen, S. V., & Ashton, M. C. (2001). Big five factors and facets and the prediction of behavior. *Journal of Personality and Social Psychology*, 81(3), 524–539.

R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rammstedt, B., Goldberg, L. R., & Borg, I. (2010). The measurement equivalence of big-five factor markers for persons with different levels of education. *Journal of Research in Personality*, 44(1), 53–61.

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin*, 132(1), 1–25.

Saucier, G., & Ostendorf, F. (1999). Hierarchical subcomponents of the big five personality factors: a cross-language replication. *Journal of Personality and Social Psychology*, 76(4), 613–627.

Scherpenzeel, A. (2011). Data collection in a probability-based internet panel: How the LISS panel was built and how it can be used. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 109(1), 56–61.

Scherpenzeel, A., & Das, J. W. M. (2010). "True" longitudinal and probability-based internet panels. In J. W. M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and Behavioral Research and the Internet* (pp. 77–103). London: Taylor & Francis.

Schmitt, D. P., Realo, A., Voracek, M., & Allik, J. (2008). Why can't a man be more like a woman? Sex differences in Big Five personality traits across 55 cultures. *Journal of Personality and Social Psychology*, 94(1), 168–182.

Schwaba, T., Luhmann, M., Denissen, J. J. A., Chung, J. M., & Bleidorn, W. (2018). Openness to experience and culture-openness transactions across the lifespan. *Journal of Personality and Social Psychology*, 115, 118–136.

Soto, C. J., & John, O. P. (2017a). Short and extra-short forms of the Big Five Inventory–2: The BFI-2-S and BFI-2-XS. *Journal of Research in Personality*, 68, 69–81.

Soto, C. J., & John, O. P. (2017b). The next big five inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, 113(1), 117–143.

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2011). Age differences in personality traits from 10 to 65: big five domains and facets in a large cross-sectional sample. *Journal of Personality and Social Psychology*, 100(2), 330–348.

Srivastava, S., John, O. P., Gosling, S. D., & Potter, J. (2003). Development of personality in early and middle adulthood: Set like plaster or persistent change?. *Journal of Personality and Social Psychology*, 84(5), 1041–1053.

Terracciano, A., Sutin, A. R., McCrae, R. R., Deiana, B., Ferrucci, L., Schlessinger, D., … Costa, P. T. (2009). Facets of personality linked to underweight and overweight. *Psychosomatic Medicine*, 71(6), 682–689.

Weisberg, Y. J., DeYoung, C. G., & Hirsh, J. B. (2011). Gender differences in personality across the ten aspects of the big five. *Frontiers in Psychology*, 2, 178