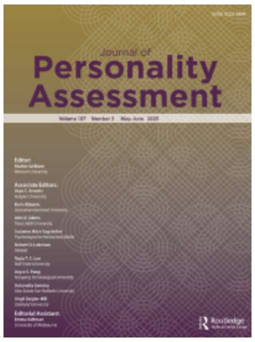
Developing and improving personality inventories using generative artificial intelligence: the psychometric properties of a short HEXACO scale developed using ChatGPT 4.0

Barends, A.J.; Vries, R.E. de



Journal of Personality Assessment

Routledge
Taylor & Francis Group

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/hjpa20

Developing and Improving Personality Inventories Using Generative Artificial Intelligence: The Psychometric Properties of a Short HEXACO Scale Developed Using ChatGPT 4.0

Ard J. Barends & Reinout E. de Vries

To cite this article: Ard J. Barends & Reinout E. de Vries (27 Dec 2024): Developing and Improving Personality Inventories Using Generative Artificial Intelligence: The Psychometric Properties of a Short HEXACO Scale Developed Using ChatGPT 4.0, Journal of Personality Assessment, DOI: 10.1080/00223891.2024.2444454

To link to this article: https://doi.org/10.1080/00223891.2024.2444454

View supplementary material

Published online: 27 Dec 2024.

Submit your article to this journal

Article views: 918

View related articles

View Crossmark data

JOURNAL OF PERSONALITY ASSESSMENT
https://doi.org/10.1080/00223891.2024.2444454

Routledge
Taylor & Francis Group

OPEN ACCESS

Check for updates

Developing and Improving Personality Inventories Using Generative Artificial Intelligence: The Psychometric Properties of a Short HEXACO Scale Developed Using ChatGPT 4.0

Ard J. Barends[1] and Reinout E. de Vries[2]

[1]Institute for Criminal Law and Criminology, Leiden, Leiden University, Leiden, The Netherlands; [2]Department of Experimental and Applied Psychology, Institute for Brain and Behavior Amsterdam, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

ABSTRACT

In the current study, we investigated the utility of generative AI for survey development and improvement. To do so, we generated a 24-item HEXACO personality inventory using ChatGPT 4.0, the ChatGPT HEXACO inventory (CHI), and investigated whether ChatGPT could modify the CHI to either improve its internal consistency or its content validity. Additionally, we compared the psychometric properties of the different versions of the CHI to a conceptually similar short personality inventory. Specifically, we compared the three CHI versions with the Brief HEXACO inventory (BHI) in terms of their alpha reliabilities and their convergent and discriminant correlations with the HEXACO-60 and criterion-related validity with authoritarianism and social dominance orientation. Participants (*N* = 682) completed the BHI and HEXACO-60 and were randomly assigned to complete one of the three CHI versions. The results showed generally comparable psychometric properties of the three CHI versions and the BHI. However, ChatGPT could not improve specific psychometric properties of the CHI. That is, although the results show promise for the use of ChatGPT in developing questionnaires, it may not offer a shortcut to further improve specific psychometric properties.

ARTICLE HISTORY
Received 23 September 2024
Revised 2 December 2024

The generative Artificial Intelligence tool ChatGPT is a web-based chatbot that allows users to automatically generate text by typing commands, so-called prompts (Wu et al., 2023). Since the public release of ChatGPT in 2022, it has resulted in scientists investigating various applications. For instance, it has been used to conduct a literature review (Haman & Školník, 2024), to infer personality traits from written text and video interviews (Derner et al., 2024; Zhang et al., 2024), or to generate personas to complete personality inventories (de Winter et al., 2024). However, so far, we are not aware of any research that has investigated how generative AI can help writing and refining questionnaire items. Specifically, in this project, we investigated the psychometric properties of a short HEXACO personality inventory generated using ChatGPT 4.0.

Short personality inventories often consist of between one and five items per personality trait (De Vries, 2013; cf. Credé et al., 2012). Researchers develop short personality scales to serve a more limited set of functions than full length personality scales (Ziegler et al., 2014). Their application is particularly useful for large scale research projects that otherwise would not include a personality measure at all, for instance, in the world values survey (e.g., Ludeke & Larsen, 2017). Short personality scales can be quickly completed by participants; however, they have some drawbacks. Specifically, with just a few items per scale to measure broad personality traits, there is a tradeoff between internal consistency and content validity (e.g., Credé et al., 2012; Ziegler et al., 2014). Specifically, most longer scales include items that cover similar content in somewhat different wording, thereby increasing internal consistency (Smith et al., 2000). In a similar vein, when constructing a short (few items) scale, researchers can choose to retain a few relatively tautological items to create an internally consistent but narrow measure of the intended construct (a so-called bloated specific). However, with short scales, it is better practice to avoid repeating similar content to more optimally cover the broad content of the original personality trait space. In this way, while sacrificing internal consistency, researchers may be able to increase content—and thus convergent and predictive—validity. Optimally, within the boundaries of short scales, researchers would like to optimize both internal consistency and content validity. However, so far, experts have not yet established any specific guidelines or rules that may help changing the psychometric properties by other means than by increasing or decreasing the extent to which items are tautological to each other.

CONTACT Ard J. Barends ✉ a.j.barends@law.leidenuniv.nl Institute for Criminal Law and Criminology, Leiden, Leiden University, Steenschuur 25, Leiden 2311 ES, The Netherlands.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/00223891.2024.2444454.

The goal of the current exploratory study was to investigate how well ChatGPT 4.0 can generate a short HEXACO personality inventory to measure its six underlying personality traits: Honesty-humility, Emotionality, eXtraversion, Agreeableness, Conscientiousness, and Openness to experience (Ashton & Lee, 2020). We compared the psychometric properties of this ChatGPT generated HEXACO personality questionnaire to a previously validated Brief HEXACO Inventory (BHI; De Vries, 2013; Julian et al., 2022). Critically, we compared the ChatGPT HEXACO Inventory (CHI) and BHI in terms of their internal consistency and their convergent and discriminant validity with the 60-item HEXACO Personality Inventory (Ashton & Lee, 2009). Finally, we also investigated the criterion-related validity in relation to authoritarianism and Social Dominance Orientation (SDO). Prior research has demonstrated that authoritarianism and SDO are mainly related to respectively openness to experience and honesty-humility (e.g., De Vries et al., 2022; Lee et al., 2010; Leone et al., 2012). These comparisons allow us to check how questionnaires generated by ChatGPT hold up when compared to a questionnaire developed by a human (i.e., the BHI). Moreover, we tested whether ChatGPT could also be used to improve upon specific psychometric properties (internal consistency reliability and content validity) if instructed to modify the CHI to optimize one of these psychometric properties. If successful, this allows researchers to more efficiently adapt (short) questionnaires to optimize specific psychometric properties.

Methods

Sample

As part of an undergraduate methods course, students recruited participants for the study within their personal network. In total, 693 respondents completed the study. We checked for noncompliant responding by analyzing the response patterns on the HEXACO-60 using the procedure developed by Lee and Ashton (2018) and validated by Barends and de Vries (2019). Moreover, we also used the speed per item procedure by Wood et al. (2017) and dropped responses if a respondent on average took less than 1 s per HEXACO-60 item. After excluding 11 noncompliant responses based on these two checks, the final sample used in the study consisted of 682 participants (223 men, 457 women, 2 other, M_{age} = 35.28 years, $SD = 17.66$).

Procedure

At first, all participants completed demographic questions, the HEXACO-60, and the BHI. Subsequently, they were randomly assigned to complete one out of three short HEXACO inventories generated using ChatGPT 4.0 (see materials). Each version of the CHI was completed by between 212 and 242 respondents. Finally, they completed the authoritarianism and SDO measures.[1]

Materials

HEXACO-60

All participants completed the Dutch version of the HEXACO-60 (Ashton & Lee, 2009; De Vries et al., 2009). This questionnaire measures each HEXACO trait with ten items per domain on a five-point Likert scale (1 = strongly disagree; 5 = strongly agree). Reliabilities for the HEXACO-60 in the full sample ranged from .74 to .81 (see the supplemental files for details).

BHI

All participants completed the Dutch version of the Brief HEXACO Inventory (BHI; De Vries, 2013). The items of the BHI and HEXACO-60 do not overlap as the BHI was written based on a simplified version of the HEXACO-PI-R, the HEXACO Simplified Personality Inventory (HEXACO-SPI; see for instance De Vries et al., 2020). The BHI measures each HEXACO trait with four items per domain on a five-point Likert scale (1 = strongly disagree; 5 = strongly agree). The reliabilities ranged from .45 to .61 in the full sample.

CHI versions

The baseline version of the CHI (CHI-B) was generated using ChatGPT 4.0 and was completed by 228 participants. To generate the items, we first provided ChatGPT 4.0 with the English language definitions of the HEXACO domains and facets (four facets per domain) from the hexaco.org website. We subsequently prompted ChatGPT to generate 24 Likert style items to measure HEXACO personality, with four items per HEXACO domain. However, we also gave specific instructions to ChatGPT to use several rules regarding item generation (Clark & Watson, 2019; De Vries et al., 2016). Specifically, ChatGPT had to create items that (1) were not tautological and (2) did not include colloquialisms or slang. Furthermore, the scale needed to (3) consist of single-barreled statements and (4) have at least one out of every four items that was negatively keyed. Moreover, the items needed to be (5) observable to others, (6) as neutral as possible in terms of social desirability, and (7) ensure sufficient variance in responses between respondents (i.e., have a high standard deviation).

The items were generated in English and were manually checked whether they complied with the above criteria. In case problems were detected (in total two issues were detected), we highlighted the issue to ChatGPT and required it to try rewrite the item using the same criteria.[2] Subsequently, we requested ChatGPT to translate the items into Dutch and to stay as close as possible to the original formulation. Again, if issues with the translation were detected (in total eight issues), we instructed ChatGPT to try to translate the item again. All the items in the CHI did not directly overlap with any items in any of the HEXACO

[1] At the end, respondents also participated in a between subjects experiment regarding judicial punishments of offenders and they could voluntarily complete 10 judicial punishment decisions. However, these tasks were not of focal interest of the current study and are not reported here.

[2] ChatGPT made several mistakes with negative keying and generating double-barreled items. However, we overlooked the fact that the honesty-humility scale included one item that was incorrectly labeled as negatively keyed in all three versions. Similarly, the CHI-V also included an agreeableness-item that was incorrectly labeled as negatively keyed. Therefore, these specific scales did not include any negatively keyed items.

inventories in use. See the supplemental files for the specific prompts we used to generate the items and all items of the three CHI versions. The reliabilities of the baseline CHI version ranged between .31 and .63.

To construct the *reliability enhanced CHI* (CHI-R), we instructed ChatGPT to rewrite the CHI-B so that the internal consistency reliability would be optimized. In doing so, it needed to use the same item writing rules that were used to construct the baseline CHI. The correction and translation procedures were the same as for the baseline CHI. Note that we did not ask ChatGPT to cover all four facets per domain. However, ChatGPT did not utilize the possibility to restrict domain content coverage as each item covered a different facet. This reliability enhanced CHI-R was completed by 242 participants. The reliabilities of the CHI-R ranged between .45 and .65.

Similarly, the *content validity enhanced CHI* (CHI-V) was created by instructing ChatGPT to rewrite the CHI-B to optimize its content validity. Again, the rules regarding item generation and the correction and translation procedures were the same as for the CHI-B. As was true for the CHI-R, each of the final CHI-V items covered a different facet. The CHI-V was completed by 212 participants. The reliabilities of the CHI-V ranged between .22 and .62.

Authoritarianism

Authoritarianism was measured using the child-rearing values scale (Feldman & Stenner, 1997). In four statements, respondents had to indicate which out of two qualities they found more important in a child, with one statement reflecting an authoritarian value and one statement reflecting a non-authoritarian value (e.g., independence or respect for elders). The reliability was .48 in the full sample.

SDO

Social dominance orientation was measured using an eight-item scale (Ho et al., 2015) that was completed on a seven-point Likert scale (1 = strongly disagree; 7 = strongly agree). The reliability of SDO in the full sample was .82.

Ethics

The current study was waived for the requirement for approval given the non-sensitive nature of the study and the requirement that all researchers within the study have to adhere to the rules and regulations of the faculty. In line with these regulations, all respondents provided informed consent before the start of the study and were debriefed at the end of the study.

Results

Comparing the CHI-B and the BHI

We first compared the psychometric properties of the CHI-B and the BHI. We used the False Discovery Rate (Benjamini & Hochberg, 1995) to correct the p-values for multiple comparisons and took the correlations between the CHI-B and BHI scales into account as the results reflect pairwise comparisons within the CHI-B sample ($n = 228$). In line with De Vries (2013), the alpha reliabilities of the CHI-B and BHI were generally low, with respective average alpha reliabilities of .51 and .53 (see Table 1). We compared the alpha reliabilities using the procedure of Diedenhoven and Musch (2016). The results showed no significant differences between alpha reliabilities of the CHI-B and BHI scales. Similarly, the average alpha reliabilities did not significantly differ between the two inventories ($\chi^2(1) = .09$, $p = .902$).

Second, there was evidence for significant convergent correlations of the CHI-B with all six corresponding HEXACO-60 scales. To calculate the average correlation, the correlations were z-transformed and then averaged before being back-transformed into correlation coefficients. The average convergent correlation of the CHI-B was $r = .68$ ($p < .001$) and the average convergent correlation of the BHI was $r = .72$ ($p < .001$). All convergent correlations were compared using z-tests. Only the honesty-humility scale had a significantly lower convergent validity with the CHI-B than with the BHI ($z = 2.90$, $p = .026$). However, the average convergent validity did not significantly differ between the two inventories ($z = .93$, $p = .492$).

Third, to check for differences in discriminant validity, we first calculated absolute discriminant correlations to avoid averaging out positive and negative correlations. Subsequently, the absolute average correlations with the noncorresponding HEXACO-60 scales were calculated using the same procedure used to calculate the average convergent correlations. The CHI-B generally showed discriminant validity with the HEXACO-60 scales, with 60% (18 out of 30) of the noncorresponding correlations being non-significant (see Table S2). None of the discriminant correlations significantly differed between the inventories. The average of all noncorresponding correlations between the CHI-B and HEXACO-60 scales was nonsignificant ($r = .12$, $p = .071$) as was the case for the noncorresponding correlations between the BHI and HEXACO-60 scales ($r = .09$, $p = .176$). These average absolute discriminant correlations did not differ significantly from each other ($z = .43$, $p = .679$).

Finally, to compare the criterion-related validities, Table 1 shows criterion-related validities for BHI openness to experience with authoritarianism ($r = -0.29$, $p < .001$) and BHI honesty-humility with SDO ($r = -0.35$, $p < .001$) that are aligned with findings in prior research (De Vries et al., 2022; Lee et al., 2010; Leone et al., 2012). Descriptively lower, but not-significantly different criterion-related validities were found for CHI-B openness to experience with authoritarianism ($r = -0.20$, $p = .002$, $z = 1.68$, $p = .243$) and CHI-B honesty-humility with SDO ($r = -0.24$, $p < .001$, $z = 1.75$, $p = .282$). When comparing all other criterion-related correlations, including the absolute averages using the same procedures that we used for the discriminant validities, the results did not show any significant differences. Specifically, the absolute average criterion-related validity of the CHI-B with authoritarianism ($r = .09$, $p = .176$) did not significantly differ from the absolute average criterion-related correlation of the BHI with authoritarianism ($r = .15$, $p = .023$, $z = 1.06$, $p = .453$). Similarly, the absolute average

Table 1. Pairwise within sample comparisons between the Brief HEXACO (BHI) and three ChatGPT 4.0 HEXACO Inventory (CHI) versions of alpha reliabilities, convergent correlations, and (absolute) discriminant correlations with the HEXACO-60 and criterion-related correlations with authoritarianism and social dominance orientation (SDO).

	CHI-B sample ($n = 228$)														
	α			convergent r			average discriminant r			authoritarianism r			SDO r		
	BHI	CHI-B	$\chi^2(1)$	BHI	CHI-B	z	BHI	CHI-B	z	BHI	CHI-B	z	BHI	CHI-B	z
H	.38	.31	.51	.65**	.50**	2.90*	.13	.11	.30	−0.14*	.00	2.02	−0.35**	−0.24**	1.75
E	.58	.53	.79	.78**	.78**	.17	.08	.13	.86	−0.08	−0.07	.13	−0.19**	−0.16*	.71
X	.64	.63	.02	.73**	.69**	1.20	.08	.14*	1.03	−0.17*	−0.07	1.63	−0.06	.05	1.92
A	.40	.36	.16	.64**	.63**	.28	.08	.16*	1.25	−0.11	−0.13	.18	−0.11	−0.19**	1.14
C	.61	.70	4.03	.72**	.76**	1.15	.08	.07	.14	.11	.05	.99	.03	−0.05	1.26
O	.58	.56	.07	.76**	.69**	2.15	.04	.10	.89	−0.29**	−0.20**	1.68	−0.22**	−0.16*	.96
Mean	.53	.51	.09	.72**	.68**	.93	.09	.12	.43	.15*	.09	1.06	.16*	.14*	.33
	CHI-R sample ($n = 242$)														
	α			convergent r			average discriminant r			authoritarianism r			SDO r		
	BHI	CHI-R	$\chi^2(1)$	BHI	CHI-R	z	BHI	CHI-R	z	BHI	CHI-R	z	BHI	CHI-R	z
H	.50	.48	.03	.65**	.58**	1.53	.17**	.10	1.18	.07	−0.06	2.09	−0.18**	−0.36**	3.05*
E	.58	.64	1.30	.80**	.76**	1.46	.09	.10	.22	−0.10	−0.11	.26	−0.10	−0.14*	.75
X	.64	.65	.07	.73**	.72**	.24	.09	.08	.18	.03	.08	.92	−0.07	.03	1.71
A	.47	.45	.08	.66**	.66**	.07	.04	.12	1.17	−0.05	−0.08	.59	−0.12	−0.13	.08
C	.63	.59	.67	.76**	.69**	1.91	.15*	.08	1.25	.11	.02	1.55	.08	.02	1.12
O	.56	.55	.07	.74**	.53**	4.87**	.10	.13*	.54	−0.26**	−0.30**	.61	−0.22**	−0.24**	.37
Mean	.56	.56	.00	.73**	.66**	1.66	.11	.10	.23	.10	.11	.13	.13	.16	.47
	CHI-V sample ($n = 212$)														
	α			convergent r			average discriminant r			authoritarianism r			SDO r		
	BHI	CHI-V	$\chi^2(1)$	BHI	CHI-V	z	BHI	CHI-V	z	BHI	CHI-V	z	BHI	CHI-V	z
H	.48	.48	.00	.67**	.59**	1.68	.15*	.09	.92	.07	−0.04	1.52	−0.30**	−0.35**	.90
E	.50	.22	7.76*	.75**	.50**	4.96**	.11	.12	.15	.01	−0.19**	2.81*	−0.11	−0.29**	2.55
X	.64	.62	.23	.72**	.66**	1.29	.12	.07	.77	−0.05	−0.04	.15	−0.10	.01	1.74
A	.49	.57	1.14	.67**	.62**	1.09	.07	.16*	1.41	−0.07	−0.09	.21	−0.15*	−0.18**	.49
C	.57	.52	.66	.76**	.69**	1.97	.10	.06	.66	.09	.03	.84	−0.03	.04	1.07
O	.56	.61	.75	.72**	.72**	.11	.05	.11	1.14	−0.33**	−0.40**	1.23	−0.13	−0.24**	1.85
Mean	.54	.50	.29	.71**	.63**	1.98	.10	.10	.00	.10	.13	.47	.14	.19	.78

*p <.05, ** p <.01.

Notes: H: Honesty-humility; E: Emotionality; X: Extraversion; A: Agreeableness; C: Conscientiousness; O: Openness to experience; CHI-B is Baseline version of the CHI; CHI-R = the (enhanced) Reliability version; CHI-V is the (enhanced convergent) Validity version. Mean correlations reflect averages of absolute correlations.

Table 2. Statistical comparisons of the alpha reliabilities, convergent correlations, and (absolute) discriminant correlations between the three 24-item ChatGPT 4.0 HEXACO Inventory (CHI) versions.

	$\chi^2(1)$ difference of α's		z of difference convergent r's		z of difference average discriminant $\lvert r\rvert$'s	
	CHI-B vs CHI-R	CHI-B vs CHI-V	CHI-B vs CHI-R	CHI-B vs CHI-V	CHI-B vs CHI-R	CHI-B vs CHI-V
H	2.98	2.63	1.19	1.41	.09	.22
E	2.44	8.15*	.50	5.12**	.27	.54
X	.07	.05	.77	.42	.67	.72
A	.76	4.90	.55	.17	.46	.03
C	3.72	7.33*	1.51	1.62	.12	.07
O	.03	.52	2.70*	.70	.37	.18
Mean	.41	.01	.37	.90	.16	.15

*p <.05, ** p <.01.

Notes: H: Honesty-humility; E: Emotionality; X: Extraversion; A: Agreeableness; C: Conscientiousness; O: Openness to experience; CHI-B is Baseline version of the CHI ($N = 228$); CHI-R = the (enhanced) Reliability version ($N = 242$); CHI-V is the (enhanced convergent) Validity version ($N = 212$).

criterion-correlation of the CHI-B with SDO (r = .14, p = .035) did not significantly differ from the absolute average criterion-related correlation of the BHI with SDO (r = .16, p = .016, z = .33, p = .741).[3]

Comparing the CHI versions

Subsequently, we investigated whether ChatGPT could optimize internal consistency and content validity if instructed to do so (see Table 1 for the psychometric properties). Therefore, the two modified CHI versions (CHI-R and CHI-V) were compared to the CHI-B using the same procedures that was used to compare the CHI-B and the BHI, including correcting for multiple comparisons using the False Discovery Rate (Benjamini & Hochberg, 1995). However, this time, the results reflect between subject comparisons.

First, as can be seen in Table 2, there was no evidence that internal consistency reliability was improved by modifications in the CHI-R and CHI-V. Some evidence suggested that modifications significantly decreased internal consistency when optimizing content validity as two out of the six CHI-V scales had significantly lower internal consistency

[3] We also compared the inventories in terms of their correlations with gender and age. The reason is that prior research found substantial gender differences in honesty-humility and emotionality, and relations between age and honesty-humility (Moshagen et al., 2019). As there were no consistent differences, the results are reported in Table S5.

than the CHI-B. However, the average α's were .56 for the CHI-R and .50 for the CHI-V and both did not significantly differ from the average α of .51 of the CHI-B (respectively $\chi^2(1) = .41$, $p = .734$ and $\chi^2(1) = .01$, $p = .908$).

Second, there was no evidence that the modifications improved the convergent correlations as one CHI-V scale and one CHI-R scale had significantly lower (instead of higher) convergent validity than the CHI-B. The average convergent validity correlation was $r = .66$ ($p < .001$) for the CHI-R and $r = .63$ ($p < .001$) for the CHI-V.[4] Both of these convergent correlations did not significantly differ from the average absolute convergent correlation of $r = .68$ ($p < .001$) of the CHI-B, respectively, $z = .37$ ($p = .708$) and $z = .90$ ($p = .642$). Third, there was no evidence that the modifications improved the discriminant correlations. The average absolute discriminant correlation was $r = .10$ ($p = .121$) for the CHI-R and $r = .10$ ($p = .147$) for the CHI-V. Again, these did not significantly differ from the average absolute discriminant correlation of $r = .12$ ($p = .071$) of the CHI-B, respectively, $z = .16$ ($p = .931$) and $z = .15$ ($p = .974$).

Comparing the CHI modified versions and the BHI

Finally, to compare the robustness of the comparisons of the CHI-B and BHI, we also made pairwise comparisons between the modified CHI versions (CHI-R and CHI-V) and the BHI using the same procedures (see above). As can be seen in Table 1 (in the columns with the χ^2- and z-tests), there were only two significant differences (6% of the 35 comparisons) between the CHI-R and the BHI. Specifically, the CHI-R had a significantly lower convergent correlation with openness to experience ($r = .53$, $p < .001$) than the BHI ($r = .74$, $p < .001$, $z = 4.87$, $p < .001$) and the CHI-R showed a significantly higher criterion-related correlation between CHI-R honesty-humility and SDO ($r = -0.36$, $p < .001$) than was shown between BHI honesty-humility and SDO ($r = -0.18$, $p < .001$, $z = 3.05$, $p = .016$). None of the other psychometric properties differed between the inventories.

When comparing the CHI-V with the BHI, there were three significant differences (9% of 35 comparisons). These differences were all associated with emotionality. Specifically, the reliability of emotionality was $\alpha = .22$ in the CHI-V and $\alpha = .50$ in the BHI ($\chi^2(1) = 7.76$, $p = .037$) and its convergent correlation differed (CHI-V: $r = .50$, $p < .001$; BHI: $r = .75$, $p < .001$; $z = 4.96$, $p < .001$). Third, the criterion-related correlation of emotionality was significantly higher in relation to authoritarianism for the CHI-V ($r = -0.19$, $p < .001$) than for the BHI ($r = .01$, $p = .916$, $z = 2.81$, $p = .035$).

Discussion

The current study investigated the psychometric properties of a short HEXACO personality inventory generated using ChatGPT 4.0 and whether ChatGPT could modify specific psychometric properties when instructed to do so. Therefore, three different versions of the ChatGPT HEXACO Inventory (CHI) were created and correlated with the 60-item HEXACO Personality Inventory (HEXACO-60) and with measures of authoritarianism and SDO. We compared the psychometric properties of these CHI inventories to the comparably short BHI developed by a human (De Vries, 2013).

First, the baseline version of the CHI had psychometric properties comparable to the BHI as it did not differ in internal consistency, (average) convergent validity, discriminant validity, and criterion-related validity. The only exception was that the correlation of baseline CHI was significantly weaker for honesty-humility when compared to the BHI. Given the few differences in psychometric properties, the results show promise for developing useful questionnaires using generative AI. However, at the moment, the quality may not be able to surpass a questionnaire developed by a human expert.

When comparing the different CHI versions, no consistent significant differences in their psychometric properties were found. That is, the findings suggest that ChatGPT is not able to modify the internal consistency or content validity when instructed to do so. As an additional test, we compared the two additional modified CHI versions to the BHI. The differences between the modified CHI versions and the BHI in psychometric properties were comparable to those of the baseline CHI. One reason that ChatGPT may not have been able to change the psychometric properties of the CHI may be because we also instructed it to take various rules into account for item development (Clark & Watson, 2019; De Vries et al., 2016). Therefore, the room to adapt the items may have been rather limited as most changes that could influence specific psychometric properties were already fixed. Do note, however, that we did leave room for the most obvious solution for improving internal consistency reliability, namely, covering fewer facets per personality trait and thereby increasing internal consistency (Smith et al., 2000). However, as noted in the methods, ChatGPT did not use this opportunity and covered all facets per trait. We should note that we did not set a goal to achieve any specific alpha level (e.g., $\alpha = .70$). Therefore, future research may want to check whether such specific goals may work better to increase internal consistency. Notwithstanding this potential limitation, we believe that it illustrates the key point that generative AI applications are—for now—tools instead of independently thinking entities that comprehend the task at hand.

An open question is how ChatGPT generated the items. The assumption is that it followed the prompts that we provided. However, as noted by an anonymous reviewer, ChatGPT may have included personality inventory items in its training data and used these to generate new items. When we asked ChatGPT about it, it stated it did not include any copyrighted psychometric instruments in its training data. When we asked whether it had access to personality inventory items available in the public domain, such as the International Personality Item Pool (IPIP; Goldberg et al., 2006), it responded that it was not trained on those items either. Unfortunately, we can only take these answers

[4] See Tables S1-4 for the correlation matrices per CHI version sample and the overall sample.

at face value as no access is provided to the ChatGPT training data.

When researchers want to develop questionnaires using ChatGPT, careful human oversight is necessary. As noted in footnote 2, ChatGPT made several mistakes when generating negatively keyed items. Researchers using and/or expanding upon our prompts may want to carefully check whether ChatGPT correctly and consistently follows the rules included in such prompts. Future research may also want to consider a broader range of criteria than those included in our study. We only compared the criterion-related validity of the CHI and BHI to two criteria. To be clear, the criterion-space of HEXACO personality is much broader (Zettler et al., 2020). Therefore, future research may want to explore whether these initial findings generalize across a broader set of criteria.

Overall, the results show that ChatGPT can generate useful short personality inventories with generally comparable psychometric properties to a short personality inventory developed by a personality scholar. Future research may want to investigate whether—and with what quality—ChatGPT can generate items for longer inventories and for constructs that are less commonly measured. However, ChatGPT is not a panacea for solving psychometric challenges that researchers face—and struggle—to solve, such as optimizing specific psychometric properties. Therefore, at the moment, generative AI can best be used as a helpful tool in questionnaire development but may be of relatively little use to further optimize already validated instruments. That is, generative AI shows great promise as a helpful tool in questionnaire development; the question is when—or whether ever—it will surpass humans when optimizing the content, reliability, and validity of psychological instruments.

Acknowledgments

We want to thank all students who helped with the data collection.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Data availability statement

Data is available on request.

ORCID

Ard J. Barends http://orcid.org/0000-0001-7067-4463

References

Ashton, M. C., & Lee, K. (2009). The HEXACO-60: A short measure of the major dimensions of personality. *Journal of Personality Assessment*, *91*(4), 340–345. https://doi.org/10.1080/00223890902935878

Ashton, M. C., & Lee, K. (2020). Objections to the HEXACO model of personality structure—and why those objections fail. *European Journal of Personality*, *34*(4), 492–510. https://doi.org/10.1002/per.2242

Barends, A. J., & de Vries, R. E. (2019). Noncompliant responding: Comparing exclusion criteria in MTurk personality research to improve data quality. *Personality and Individual Differences*, *143*, 84–89. https://doi.org/10.1016/j.paid.2019.02.015

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *57*(1), 289–300. https://doi.org/10.1111/j.251706161.1995.tb02031.x

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, *31*(12), 1412–1427. https://doi.org/10.1037/pas0000626

Credé, M., Harms, P., Niehorster, S., & Gaye-Valentine, A. (2012). An evaluation of the consequences of using short measures of the Big Five personality traits. *Journal of Personality and Social Psychology*, *102*(4), 874–888. https://doi.org/10.1037/a0027403

De Vries, R. E. (2013). The 2 4-item brief HEXACO Inventory (BHI). *Journal of Research in Personality*, *47*(6), 871–880. https://doi.org/10.1016/j.jrp.2013.09.003

De Vries, R. E., Ashton, M. C., & Lee, K. (2009). De zes belangrijkste persoonlijkheidsdimensies en de HEXACO persoonlijkheidsvragenlijst [The six most important personality dimensions and the HEXACO personality inventory]. *Gedrag & Organisatie*, *22*(3), 232–274. https://doi.org/10.5117/2009.022.003.004

De Vries, R. E., Pronk, J., Olthof, T., & Goossens, F. A. (2020). Getting along and/or getting ahead: Differential HEXACO personality correlates of likeability and popularity among adolescents. *European Journal of Personality*, *34*(2), 245–261. https://doi.org/10.1002/per.2243

De Vries, R. E., Realo, A., & Allik, J. (2016). Using personality item characteristics to predict single-item internal reliability, retest reliability, and self-other agreement. *European Journal of Personality*, *30*(6), 618–636. https://doi.org/10.1002/per.2083

De Vries, R. E., Wesseldijk, L. W., Karinen, A. K., Jern, P., & Tybur, J. M. (2022). Relations between HEXACO personality and ideology variables are mostly genetic in nature. *European Journal of Personality*, *36*(2), 200–217. https://doi.org/10.1177/08902070211014035

de Winter, J. C. F., Driessen, T., & Dodou, D. (2024). The use of ChatGPT for personality research: Administering questionnaires using generated personas. *Personality and Individual Differences*, *228*, 112729. https://doi.org/10.1016/j.paid.20024.112729

Derner, E., Kučera, D., Oliver, N., & Zahálka, J. (2024). Can ChatGPT read who you are? *Computers in Human Behavior: Artificial Humans*, *2*(2), 100088. https://doi.org/10.1016/j.chbah.2024.100088

Diedenhoven, B., & Musch, J. (2016). Cocron: A web interface and R package for the statistical comparison of Cronbach's Alpha coefficients. *International Journal of Internet Science*, *11*(1), 51–60.

Feldman, S., & Stenner, K. (1997). Perceived threat and authoritarianism. *Political Psychology*, *18*(4), 741–770. https://doi.org/10.1111/0162-895X.00077

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. https://doi.org/10.1016/j.jrp.2005.08.007

Haman, M., & Školník, M. (2024). Using ChatGPT to conduct a literature review. *Accountability in Research*, *31*(8), 1244–1246. https://doi.org/10.1080/08989621.2023.2185514

Ho, A. K., Sidanius, J., Kteily, N., Sheehy-Skeffington, J., Pratto, F., Henkel, K. E., Foels, R., & Stewart, A. L. (2015). The nature of social dominance orientation: Theorizing and measuring preferences for intergroup inequality using the new SDO_7 scale. *Journal of Personality and Social Psychology*, *109*(6), 1003–1028. https://doi.org/10.1037/pspi0000033

Julian, A. M., Novitsky, C., Lee, K., & Ashton, M. C. (2022). Convergent validity of three brief six-factor measures of personality. *Personality and Individual Differences*, *188*, 111436. https://doi.org/10.1016/j.paid.2021.11436

Lee, K., & Ashton, M. C. (2018). Psychometric properties of the HEXACO-100. *Assessment*, *25*(5), 543–556. https://doi.org/10.1177/1073191116659134

Lee, K., Ashton, M. C., Ogunfowora, B., Bourdage, J. S., & Shin, K. H. (2010). The personality bases of socio-political attitudes: The role of Honesty–Humility and Openness to Experience. *Journal of Research in Personality*, *44*(1), 115–119. https://doi.org/10.1016/j.jrp.2009.08.007

Leone, L., Desimoni, M., & Chirumbolo, A. (2012). HEXACO, social worldviews and socio-political attitudes: A mediation analysis.

Personality and Individual Differences, *53*(8), 995–1001. https://doi.org/10.1016/j.paid.2012.07.016

Ludeke, S., & Larsen, E. G. (2017). Problems with the big five assessment in the world values survey. *Personality and Individual Differences*, *112*, 103–105. https://doi.org/10.1016/j.paid.2017.02.042

Moshagen, M., Thielmann, I., Hilbig, B. E., & Zettler, I. (2019). Meta-analytic investigations of the HEXACO Personality Inventory (-Revised): Reliability, generalization, self-observer agreement, intercorrelations, and relations to demographic variables. *Zeitschrift Für Psychologie*, *227*(3), 186–194. https://doi.org/10.1027/2604/a000377

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment*, *12*(1), 102–111. https://doi.org/10.1037/1040-3590.12.1.102

Wood, D., Harms, P. D., Lowman, G. H., & DeSimone, J. A. (2017). Response speed and response consistency as mutually validating indicators of data quality in online samples. *Social Psychological and Personality Science*, *8*(4), 454–464. https://doi.org/10.1177/1948550617703168

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.-L., & Tang, Y. (2023). A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, *10*(5), 1122–1136. https://doi.org/10.1109/JAS.2023.123618

Zhang, T., Koutsoumpis, A., Oostrom, J. K., Holtrop, D. J., Ghassemi, S., & De Vries, R. E. (2024). Can Large Language Models assess personality from Asynchronous Video Interviews? A comprehensive evaluation of validity, reliability, fairness, and rating patterns. *IEEE Transactions on Affective Computing*, *15*(3), 1769–1785. https://doi.org/10.1109/TAFFC.2024.3374875

Zettler, I., Thielmann, I., Hilbig, B. E., & Moshagen, M. (2020). The nomological net of the HEXACO model of personality: A large-scale meta-analytic investigation. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *15*(3), 723–760. https://doi.org/10.1177/1745691619895036

Ziegler, M., Kemper, C. J., & Kruyen, P. (2014). Short scales – Five misunderstandings and ways to overcome them. *Journal of Individual Differences*, *35*(4), 185–189. https://doi.org/10.1027/1614-0001/a000148