Matching GPT-simulated Populations with Real Ones in Psychological Studies—The Case of the EPQR-A Personality Test

GREGORIO FERREIRA, JACOPO AMIDEI, RUBÉN NIETO, and ANDREAS KALTENBRUNNER, Universitat Oberta de Catalunya, Barcelona, Spain

This article analyzes how well OpenAI's LLM GPT-4 can emulate different personalities and simulate populations to answer psychological questionnaires similarly to real population samples. For this purpose, we performed different experiments with the Eysenck Personality Questionnaire-Revised Abbreviated (EPQR-A) in three different languages (Spanish, English, and Slovak). The EPQR-A measures personality on four scales: extraversion (E: sociability), neuroticism (N: emotional stability), psychoticism (P: tendency to break social rules, and not having empathy), and lying (L: social desirability).

We perform a comparative analysis of the answers of synthetic populations with those of two real population samples of Spanish students as well as the unconditioned baseline personality of GPT. Furthermore, the impact of time (what year the questionnaire is answered), questionnaire language, and student age and gender are analyzed.

To our knowledge, this is the first time the EPQR-A test has been used to assess the GPT´s personality and the impact of different language versions and time are measured.

Our analysis reveals that GPT-4 exhibits an extroverted, emotionally stable personality with low psychoticism levels and high social desirability. GPT-4 replicates some differences observed in real populations in terms of gender but only partially replicates the results for real populations.

CCS Concepts: • **Applied computing** → **Psychology**; • **Computing methodologies** → *Natural language generation*;

Additional Key Words and Phrases: Large Language Models, EPQR-A test, GPT, synthetic populations, personality test

ACM Reference format:
Gregorio Ferreira, Jacopo Amidei, Rubén Nieto, and Andreas Kaltenbrunner. 2025. Matching GPT-simulated Populations with Real Ones in Psychological Studies—The Case of the EPQR-A Personality Test. *ACM Trans. Comput. Healthcare* 6, 2, Article 26 (April 2025), 33 pages.
https://doi.org/10.1145/3712301

1 Introduction

Questionnaires and surveys are efficient research methods used for acquiring information about individuals, especially useful to unearth information that is not directly observable or measurable [39, 61]. Usually, when

Authors' Contact Information: Gregorio Ferreira, Universitat Oberta de Catalunya, Barcelona, Spain; e-mail: jferreirade@uoc.edu; Jacopo Amidei, Universitat Oberta de Catalunya, Barcelona, Spain; e-mail: jamidei@uoc.edu; Rubén Nieto, Universitat Oberta de Catalunya, Barcelona, Spain; e-mail: rnietol@uoc.edu; Andreas Kaltenbrunner (corresponding author), Universitat Oberta de Catalunya, Barcelona, Spain; e-mail: kaltenbrunner@gmail.com.

ACM 2637-8051/2025/4-ART26
https://doi.org/10.1145/3712301

health-related variables are assessed, subjective experiences are accessed that can be only self-reported by people using questionnaires. This can be for example variables such as personal characteristics—e.g., personality [6, 12, 19, 21]—quality of life [20, 28], mood [5, 66], or the perceived severity of a health problem or pain [53, 65]. To collect information about these subjective experiences, questionnaires, and surveys are useful tools. Among others, they can be useful to collect views of people about a great diversity of topics such as healthcare services utilization (or usage intention), opinions about a service or a new procedure/intervention, and/or the occurrence of a health issue for epidemiologic purposes. Given these applications, their efficiency in collecting data and their low cost, surveys, and questionnaires are largely used in the health field [61].

However, defining sound surveys and questionnaires is not reduced to providing a sequence of questions. The overall structure, flow, coherence, and adequacy of the questions have to be taken into account. The definition of surveys and questionnaires involves steps that were consolidated by years of research [58]. Apart from other necessary steps, the two most important and time-consuming ones are the pre-test (or pilot test), and the test of psychometric properties. In the pre-test, the questions are administered to a few participants to gather information which allows for revision of the survey or questionnaire. This step ideally should involve several iterations. Once the final version is reached, testing the psychometric properties comes into play. This step is needed to examine reliability (the extent to which the data collected by a survey or a questionnaire are reproducible and consistent), and validity (to what extent the survey or questionnaire assesses what it is intended to measure). These two steps are complex, time-consuming, and usually require expending a lot of resources since they involve the computation of multiple indices and applying the questionnaire to different populations [58].

Facing these difficulties, we want to examine to what extent it is possible to help researchers and/or healthcare professionals simulate populations for testing surveys or questionnaires by using **Large Language Models (LLMs)**. Along these lines, we have started with performing several tests with a questionnaire commonly used to assess personality: the **Eysenck Personality Inventory (EPQ)** [19]. It was created to assess Extroversion (E; sociability), Neuroticism (N; emotional stability), Psychoticism (P; aggressiveness, lack of empathy, and concern for social norms) and also has a scale to assess social desirability (L). We have chosen this questionnaire given its widespread use in psychology and availability in multiple languages (e.g., English [21], Spanish [23, 57], Brazilian [59], Slovak [17], French [43], Turkish [40], Greek [2], or Urdu [44]), facilitating multi-language experiments as part of our research.

Our approach involves replicating two psychological tests of the EPQ in Spanish: Sandín et al. [57] and García-González et al. [23] using synthetic populations of LLMs. We focus on the OpenAI's LLM family (specifically, GPT-3.5 turbo, GPT-4, and GPT-4 preview) and start with an assessment of GPT's personality. Accordingly, our first research question is as follows:

RQ1 *What is the personality of GPT?* To address this question, we evaluate GPT's personality across various experimental presentations and language versions of the questionnaire. We also examine different versions and temperatures of GPT.

Answering RQ1 enables us to evaluate the consistency of system responses and establish parameters for subsequent experiments. Additionally, it allows us to establish a baseline personality for our model to ascertain if the synthetic populations generated by GPT are influenced by its intrinsic personality.

Then we shift our focus to the definition of GPT synthetic populations. Consequently, our second research question is as follows:

RQ2 *Can GPT be steered towards answering like human population samples to the **Eysenck Personality Questionnaire-Revised Abbreviated (EPQR-A)** personality test?* To answer this question we simulate the population from two previous published works [23, 57] by setting the same parameters as described by the original authors in terms of age and gender composition of the test population, to simulate as best as possible their population.

RQ2 aims to investigate the feasibility of replicating the findings of Sandín et al. [57] and García-González et al. [23] using synthetic populations. We compare results across different factors like temperature and GPT version. Interestingly, during our attempts to generate GPT synthetic populations, we consistently observed a trend of positive bias wherein GPT predominantly produced positive populations, with minimal or no representation of negative human characteristics. Even when prompted to consider negative human features, GPT generated personalities that portrayed potentially negative traits from a positive perspective. In essence, our findings echo concerns raised by other researchers [1, 13, 31, 69]: LLMs populations tend to flatten the representativeness of groups.

Lastly, we examined how synthetic populations generated by GPT respond differently based on variables such as gender, age, language, and the years in which the questionnaire was administered. Therefore, our last three research questions are as follows:

RQ3 *Do the personalities generated by GPT answer differently depending on their characteristics?* We performed a fine-grained analysis of specific groups of personas making up the synthetic population (by gender and age groups) to check if they match the expected personality or if they flat out on GPT´s personality.

RQ4 *Do the results of questionnaires answered by GPT differ depending on the language of the questionnaire?* To answer this question, we sent the questionnaire in three languages (Spanish, English, and Slovak) using the same population and prompts.

RQ5 *Do the results of questionnaires answered by GPT differ depending on divers' years?* To answer this question we used the synthetic population from [57] and we prompted GPT to answer the questionnaire as if we were in 2001.

RQ3 enables us to examine whether subgroups within the synthetic population replicate the findings of the original studies [23, 57]. RQ4 assesses the cross-language consistency of the GPT synthetic population. We hypothesize that individuals proficient in Spanish, English, and Slovak will respond similarly to the questionnaire regardless of the language used, reflecting consistent personality traits across languages. RQ4 allows us to verify if the synthetic population exhibits this consistency. Lastly, RQ5 facilitates a comparison between [57] and the synthetic population generated by it, considering the timeframe of the original study. This enables us to assess if GPT can contextualize its responses based on the time period in which it is prompted to operate.

Regarding RQ1, our experiments show that GPT scored consistently high in E, and low in N and P. Scores in L were very high. These results were in general terms consistent across languages (the most important differences were in the P scale). Results from RQ2 show that GPT can create a sample of personas profiles, and answer the EPQR-A by simulating each specific persona. However, the GPT personality seems to bias the responses, since compared to prior studies with real population samples found in existing literature, the mean scores in E were higher, and the ones in N, were lower. For RQ3, we found differences by gender and age groups, suggesting that GPT tried to accommodate the results to different profiles consistently to what has been reported in the literature with real samples. In general terms, with RQ4 we found that language did not affect extensively the simulated samples result, supporting the consistency of results. Finally, regarding RQ5, there were no significant differences when instructing GPT to answer the questionnaire in different years.

To the best of our knowledge, the main contributions of this paper are as follows:

—We apply for the first time the EPQR-A personality test for measuring GPT personality;
—We introduce a prompting strategy that uses GPT to reproduce two previous works [23, 57] by generating a synthetic population based on their population statistics;
—We measure the stability of our results by performing a consistent analysis of the synthetic population answers across time, languages, genders and ages.

2 Related Work

The recent surge in employing psychology methodologies within the LLMs framework (see for example [32, 42]) has led Hagendorff [29] to coin the term *machine psychology*. Similarly, Pellert et al. [52] introduces the term *AI Psychometrics*. According to Hagendorff, machine psychology seeks to uncover emergent abilities in LLMs that traditional natural language processing benchmarks cannot detect. Thus, the term aims to encompass various approaches that utilize psychological methods to analyze LLMs' behavior under one umbrella.

Among the various directions that contribute to the development of machine psychology, a popular one is studying the personality of LLMs. For example, the Big Five factors [15] were used, among others, by [36, 41, 46, 52, 63] to quantify the personality traits of LLMs. Similarly, IPIP-NEO [25] was used in [63] and Short Dark Tetrad) [51] was used in [52]. In a slightly different fashion, Griffin et al. [27] investigates LLM´s behavioral profile in a *dynamic* context instead of a *static* one. While the outcomes of the aforementioned studies may vary depending on the LLMs and questionnaires used, there is enough support to draw the promising and optimistic conclusion that personality assessments for LLMs are valid and reliable. These findings hold significance, considering that personality tests are tailored for humans, and there is no guarantee beforehand that they will yield valid and reliable results for LLMs.

Alongside personality traits also value orientations [52], moral beliefs [3, 50, 52, 60], and legal reasoning [3] are been evaluated. For example, the 57-item Revised Portrait Values Questionnaire [62] was used in [52] for studying LLMs' value orientations, whereas the Moral Foundations Questionnaire [26] was used by [3, 50, 52] for studying LLMs' moral norms. Bonagiri et al. [7] introduce an information-theoretic measure (SaGE) to evaluate moral consistency in LLMs. Differently, Hagendorff et al. [30] use the cognitive reflection test [22] and the semantic illusions [18], to study the ability of LLMs to reason and make decisions. Interestingly, models such as GPT-3.5 and GPT-4 outperform humans in their ability to avoid traps embedded in the tasks.

Another type of study brings a different perspective to the area of machine psychology and aims at using LLMs to inform psychology. For example, Rao et al. [54] use the Myers–Briggs Type Indicator [47] tests to investigate the ability of LLMs to assess human personalities and conclude that this would be feasible.

A further endeavor, pursued by researchers such as [37, 41, 63], involves exploring the potential for adjusting LLMs' personalities. The objective is to study whether tailored prompts can push LLMs to replicate human personality traits. Encouragingly, findings suggest that LLMs can indeed be molded to imitate particular personality profiles.

A more general direction has brought researchers to utilize LLMs to create simulated humans, serving as experimental participants, survey respondents, or other agents. Simulated populations can streamline experiments, reducing time and costs, and can be utilized in studies unsuitable for human involvement. This concept, referred to by various names such as *guinea pigbots* [35], *silicon samples*, or *homo silicus* [34], has found application across various domains within social science. For example, the employment of LLMs as substitutes for human participants was studied in psychological research [16, 35, 50], political polling [55], software engineering research [24], teaching research [45], economics [34], social media platforms design [49, 68], market research to understand consumer preferences [8], and more generally social science research [4]. These studies yielded mixed results. While some outcomes closely mirrored the behaviors observed in real human counterparts, other research raised questions about the suitability of replacing human participants with LLMs in various social science contexts (for example, [50, 55]).

Alongside this large amount of studies that pinpoint the desirability of replacing human populations with LLMs synthetic populations, some researchers raised doubts about it [1, 13, 31, 69]. For example, Crockett and Messeri [13] underline the problem of population generalizability. Given that LLMs are trained over datasets mainly representative of young people from oversampling **Western, Educated, Industrialized, Rich and Democratic (WEIRD)** countries. This can raise the problem that the LLMs' population are mainly WEIRD-oriented and not generally representative. Similarly, Harding et al. [31] move a critic about the fact that LLMs' population

representativeness must be carefully circumscribed. Indeed Wang et al. [69] show that LLMs will misportray and flatten demographic groups due to intrinsic models' limitations. Furthermore, Wang et al. argues that the limitations of LLMs in representing demographic identities are due to the textual data format they are trained on and the loss functions employed during training. Finally, the authors of [69] conclude that LLMs, due to structural reasons, are unable to replicate the diversity of human responses, as they can represent populations as one-dimensional and lacking depth.

Additionally, Agnew et al. [1] shed light on two dimensions of potential obstacles when utilizing LLMs to simulate human behavior. The first dimension encompasses practical hurdles, including challenges in representing minority groups, the current inability of LLMs to emulate human cognition and decision-making accurately, the reliance of psychology research on various non-linguistic cues to study human cognition and behaviour, and the phenomenon of "value lock-in". This latter concept refers to LLMs reflecting attitudes from the time of their training, thus potentially limiting their ability to capture only the social and cultural conditions of that period. While advancements in technology may mitigate these practical challenges and potentially overcome them altogether, other more intrinsic obstacles could emerge. More precisely, the representation, inclusion and empowerment of participants. Replacing human participants with AI may disrupt the intersubjectivity between researcher and participant, where, in Agnew et al. [1] words, intersubjectivity "can be summarized simply as the researcher's assumption that if they were to take another person's position, they would see or experience the situation as that person does". The authors of [1] conclude their study by warning us about the risk of anthropomorphizing LLMs.

In this article, we investigate the feasibility of utilizing LLMs as a substitute for human populations in healthcare questionnaires and surveys. To our knowledge, we are the first to evaluate the personality traits of GPT using the EPQR-A personality test and to assess its consistency across different years, languages (more precisely, Spanish, English, and Slovak), genders, and ages.

3 Methods

We used for our experiments the Eysenck Personality Questionnaire-Revised (EPQR-A), which is an abbreviated version of the Eysenck Personality Inventory [19], containing 24 items for assessing 4 different scales (6 items each): *Extraversion (E), Neuroticism (N), Psychoticism (P), Lie (L)* [21]. Each item has a dichotomous response (yes or no), and a score for each scale can be computed by summing individual items (resulting in a range from 0 to 6). The EPQR-A was originally tested in English obtaining relevant results [21], and has been validated in many different languages such as Spanish [57], Brazilian [59], Slovak [17], French [43], Turkish [40], Greek [2], or Urdu [44]. This made this tool an excellent one for our objectives.

3.1 Types of Experiments

Using the EPQR-A questionnaire we performed the following two experiments with different versions of GPT. In all cases we used English instructions, but provided the questionnaire in three different languages: English, Spanish, and Slovak. The rationale for using this setting is that GPT's responses were more accurate and consistent when instructions were given in English, even if the questionnaire was in a different language. A qualitative analysis of the answers revealed that, with English instructions, the models' responses adhered more closely to the expected structure (given the instructions) and context of the questions. For example, answers different from yes and no or answers adding non-sensical characters happened occasionally when non-English instructions were used. Our findings align with recent research, such as [33, 48, 64].

(1) Testing the *personality of GPT* and the consistency of the EPQR-A scores. We report means and standard deviations of 20 runs to measure potential variability in the answers. We tested three models: gpt-3.5-turbo-0125, gpt-4-0613, and gpt-4-0125-preview with four different temperature settings: 0, 0.5, 1, and 1.5. The temperature parameter controls the randomness in the responses of LLMs. Temperature 0 indicates

minimal randomness, while temperature 2 indicates maximum randomness. Temperature 1, the default value for GP, represents an average level of randomness.

(2) Testing the properties of the EPQR-A when administered to a *sample population generated by GPT*-4: we generate two simulated samples of students equivalent in size, age, and gender composition as the one used in two Spanish validation studies: Sandín et al. [57] and García-González et al. [23]. The first consisted of 199 students and the second of 826. We chose these studies because they provided the most details about their sample population.

3.2 LLM Prompting Strategy

To obtain sample populations from GPT-4, we performed the following prompt strategy. We first calculated fixed lists of `id`, `gender`, and `age` reflecting the corresponding statistics from the real-world samples used in Sandín et al. [57] (149 females aged 21.5 ± 6.3, 50 males aged 22.1 ± 7.1) and García-González et al. [23] (655 females aged 18.9 ± 1.56, 171 males aged 19.6 ± 1.99) using Python's statistics library. Then this list was passed to GPT-4, iteratively with instructions to generate a description of a personality, which led to personality descriptions like the one listed below:

```
Female, 19: A psychology major interested in developmental psychology. Her empathy and understanding make her a great peer counsellor, but she sometimes becomes overly involved in the personal issues of others.
```

Apart from the age and gender distributions the only difference between the prompts used for Sandín et al. [57] and García-González et al. [23] are in the origin of the students which are either "the Universidad Autonoma de Madrid (UAM)" or "from different universities in Spain". The specific detailed prompts are explained in Appendix B1.

Each profile is unique, with distinct combinations of the parameters to reflect the diversity of the student body. Once the virtual population sample had been created we sent the EPQR-A personality test questionnaires in the different languages to GPT using OpenAI's API with the task of answering them by impersonating the 199 (or 826) generated personalities one by one. In more detail, we first specify a personality as a system role in the API call followed by the prompt to answer the questionnaire. See Appendix B2 for the specific instructions used.

3.3 Prompt Development

When crafting the prompts to generate the virtual populations, we adopted an iterative approach and tested different configurations. In the initial iterations, we observed a *positive bias*, where the generated descriptions systematically leaned toward positive attributes and outcomes, underestimating negative ones. To mitigate this bias, we employed a trial-and-error strategy over four iterations. In the first iteration, the positive bias effect emerged clearly. Consequently, we adjusted the instructions, adding specific attributes in the second iteration. This led to an improved trait spectrum in the personas but still exhibited a bias towards positive traits. For the third iteration, we refined the instructions further to include more negative traits, aiming to balance the generated personas. This reduced the initial positive bias, although the descriptions still showed some skewing towards higher social desirability. In response, we conducted a fourth iteration, which helped further reduce the social desirability bias. More details about our iterative process can be found in Appendix A1. While we did not use formal prompt engineering, our iterative approach to refining instructions based on a trial-and-error strategy proved effective in improving the generation of virtual personas. Our findings suggest that using GPT to create virtual populations that accurately mimic "real-world" ones requires continuous refinement and adjustment of the GPT instructions. In conclusion, to construct statistically meaningful and realistic populations for psychological research, researchers need to tailor their GPT instructions to overcome the limitations imposed by GPT's guardrails.

Table 1. GPT Personality Mean (± sd) of the EPQR-A Test in Spanish for Different Temperatures

Temperature	Scale	gpt-3.5-turbo-0125	gpt-4-0613	gpt-4-0125-preview
0	E	6.00 (± 0.00)	5.95 (± 0.22)	6.00 (± 0.00)
	N	0.00 (± 0.00)	0.10 (± 0.45)	0.00 (± 0.00)
	P	2.00 (± 0.00)	2.00 (± 0.00)	2.00 (± 0.00)
	L	1.60 (± 0.50)	1.25 (± 0.44)	1.45 (± 0.51)
0.5	E	5.75 (± 0.55)	5.60 (± 0.60)	5.90 (± 0.31)
	N	0.15 (± 0.49)	0.40 (± 0.68)	0.30 (± 0.92)
	P	1.90 (± 0.31)	1.90 (± 0.31)	2.00 (± 0.00)
	L	1.55 (± 0.51)	1.15 (± 0.75)	1.25 (± 0.55)
1	E	5.15 (± 0.99)	5.80 (± 0.41)	5.30 (± 0.80)
	N	0.80 (± 0.89)	0.55 (± 0.76)	0.40 (± 0.60)
	P	2.00 (± 0.56)	1.85 (± 0.37)	1.60 (± 0.75)
	L	1.15 (± 0.49)	1.25 (± 0.64)	1.35 (± 0.49)
1.5	E	5.00 (± 0.79)	5.25 (± 0.85)	5.29 (± 0.85)
	N	1.30 (± 1.72)	1.85 (± 1.53)	0.95 (± 1.24)
	P	2.00 (± 0.79)	1.55 (± 0.69)	1.81 (± 0.51)
	L	1.80 (± 0.83)	1.45 (± 0.94)	1.43 (± 0.75)

20 iterations per model.

3.4 Postprocessing of the Answers

With the answers from the population of virtual personas, we compute descriptive statistics for each of the 4 scales of the EPQR-A, and we tested reliability by computing Cronbach's α [14] values. This is an index frequently used to evaluate the internal consistency of a set of items [67]. Cronbach's α is a way of assessing reliability by comparing the amount of shared variance, or covariance, among the items making up a scale to the amount of overall variance. The idea is that if the scale is reliable, there should be a great deal of covariance among the items relative to the variance [11]. Cronbach's α is considered poor if it is below 0.70; fair when it is between 0.70 and 0.79; good when it is between 0.80 and 0.89; and excellent when it is above 0.90 [10].

4 Results

4.1 RQ1: Testing the Personality of GPT

We start with analyzing the response of GPT to the EPQR-A test. We first analyze the impact of different temperatures on the responses of three different versions of GPT. As can be seen in Table 1, in all settings GPT scores high in E, but low in N, P and L. This would imply that it adopts a personality characterized by extroversion, emotional stability, and trying to follow social norms (with high levels of desirability). As expected setting GPTs temperature to 0 or 0.5 leads to less variability in the answers (lower standard deviation).[1] The differences between a parameter setting of 1 or 1.5 are less clear, and we have thus opted to use the default setting of GPT for the temperature (i.e., a temperature of 1) in the remainder of this article.

Two different versions of GPT-4 were considered in our tests: GPT-4-0613 and GPT-4-0125-preview. After evaluating the results, we concluded that both models depict the same personality as measured through the

[1]We also performed experiments with setting the temperature to 2, but then in most of the 20 iterations GPT did not correctly answer the questionnaire and we have omitted these results.

Table 2. GPT Personality Mean (± sd) for Different Languages

Language	Scale	gpt-3.5-turbo-0125	gpt-4-0613	gpt-4-0125-preview
Spanish	E	5.24 (± 1.42)	5.15 (± 1.64)	5.63 (± 0.72)
	N	0.24 (± 0.65)	0.38 (± 1.02)	0.27 (± 0.92)
	P	0.73 (± 0.95)	0.90 (± 0.89)	0.88 (± 0.83)
	L	1.04 (± 0.24)	1.08 (± 0.34)	1.07 (± 0.26)
English	E	4.87 (± 2.02)‡	5.03 (± 2.03)‡	5.65 (± 1.32)
	N	0.83 (± 1.41)	1.33 (± 1.76)	1.14 (± 1.63)
	P	1.21 (± 1.03)	1.09 (± 0.96)	0.89 (± 0.91)
	L	5.89 (± 0.65)	5.73 (± 0.79)	5.67 (± 0.89)
Slovak	E	3.64 (± 2.74)∗	3.66 (± 2.72)∗	4.51 (± 2.38)‡
	N	0.86 (± 1.44)	1.24 (± 1.78)	0.75 (± 1.35)
	P	1.24 (± 0.55)	1.38 (± 0.72)∗	1.40 (± 0.67)∗
	L	5.52 (± 1.05)	5.55 (± 0.94)	5.41 (± 1.30)

Temperature 1 and 100 iterations per model. Cases where the results are not significantly different from the result of real populations of Sandín et al. [58] are indicated with ∗ and for García-González et al. [24] with ‡ (i.e. $p \geq 0.01$ in a t-test). Detailed results of the t-tests in Tables C2 (Spanish), C3 (English), and C4 (Slovak) in the Appendix.

EPQR-A test. For our research, we decided to use GPT-4-0125-preview, as it gave better performance and the ability to obtain the same structured answer across the different experiments.

When analyzing the impact of using different languages for the questionnaire we find (results shown in Table 2) that the results are very similar across the different languages in the N and L scales (there were no significant differences in mean scores). The difference between Spanish and English/Slovak in the L scale can be explained by the change in the scoring introduced by Sandín et al. [57], who reversed the L scores. So lower scores indicate greater social desirability in the Spanish version of the EPQR-A questionnaire, while the same holds for the higher scores in English and Slovak. Furthermore, it is interesting to observe the lower scores in E for Slovak together with higher P scores.

4.2 RQ2: Testing the EPQR-A When Administered to Two Different Virtual Populations

Next, we analyze how GPT answers the questionnaire when simulating human populations. We perform this analysis for two different real populations of Spanish students used in the studies of Sandín et al. [57] and García-González et al. [23].

The first four rows (grouped under Population-type total) of Table 3 contain descriptive statistics of the EPQR-A questionnaire results for the real sample populations from these studies as well as those of the simulated samples with GPT. For simplicity, in the remainder of this article, we will only report results for GPT-4-0125-preview. However, Table B1 in the Appendix contains the results of the same experiments for GPT-3.5 (the results for the two versions of GPT are very similar in most cases). The table contains two scores for simulated samples for Sandín et al. [57] (Sim. and Sim. 2001). For now, we will only analyze the Sim. column, the other one corresponds to experiments where GPT was asked to answer like it were 2001 and will be discussed in Section 4.5.

At the descriptive level, when comparing the real samples we observe that García-González et al. [23] reported greater scores in E, and lower scores in N than the one of Sandín et al. [57]. The differences to the simulated sample for García-González et al. [23] are all statistically significant and in the following directions: the simulated sample displayed greater scores in E, P and L, and slightly lower ones in N than the real one. The same pattern was observed when comparing the simulated sample for Sandín et al. [57] with the corresponding real sample and also here paired T-tests report significant differences for all the scales.

Table 3. Mean (± sd) Scores for the Real and Simulated Samples for Spanish

Mod.	Population Type / Size	Scale	Mean (± sd)				
			Sandín [58]			Garcia [24]	
			Real	Sim.	Sim. 2001	Real	Sim.
gpt-4-0125-preview	Total Ref. [58] = 199 Ref. [24] = 826	E	3.05 (± 1.67)	5.35 (± 1.69)	5.29 (± 1.74)‡	4.56 (± 1.83)	5.54 (± 1.36)
		N	3.75 (± 1.86)	1.93 (± 1.85)	2.16 (± 1.89)‡	2.52 (± 1.85)	1.50 (± 1.57)
		P	1.59 (± 1.20)	3.34 (± 1.26)	3.56 (± 1.29)‡	1.71 (± 1.20)	3.36 (± 1.24)
		L	3.20 (± 1.61)	3.86 (± 1.48)	3.78 (± 1.44)‡	3.29 (± 1.61)	3.89 (± 1.37)
	Male Ref. [58] = 50 Ref. [24] = 171	E	2.70 (± 1.56)	5.48 (± 1.28)	5.54 (± 1.37)	4.62 (± 1.70)	5.61 (± 1.28)
		N	3.70 (± 1.78)	**1.06 (± 1.46)**	**1.58 (± 1.70)**	**2.06 (± 1.76)**	**1.11 (± 1.38)**
		P	1.71 (± 1.31)	**3.82 (± 1.47)**	3.94 (± 1.33)	**1.88 (± 1.25)**	**3.70 (± 1.25)**
		L	**3.75 (± 1.62)**	**4.38 (± 1.07)**∗	**4.44 (± 0.97)**∗	**3.84 (± 1.56)**	**2.98 (± 1.55)**
	Female Ref. [58] = 149 Ref. [24] = 655	E	3.17 (± 1.69)	5.30 (± 1.80)	5.20 (± 1.85)	4.55 (± 1.87)	5.52 (± 1.38)
		N	3.77 (± 1.89)	**2.23 (± 1.88)**	**2.36 (± 1.92)**	**2.64 (± 1.86)**	**1.60 (± 1.60)**
		P	1.55 (± 1.16)	**3.18 (± 1.15)**	3.43 (± 1.25)	**1.25 (± 1.67)**	**3.27 (± 1.22)**
		L	**3.01 (± 1.57)**	**3.69 (± 1.55)**	**3.56 (± 1.50)**	**3.15 (± 1.59)**	**4.13 (± 1.21)**

∗ indicates the cases where the real population is not significantly different from the simulated one (Tables C5 and C9), while ‡ the cases where the 2001 population is not significantly different from the undated population of Sandín et al. [58] (Table C6). Underlined values indicate cases where the values are not significantly different from the GPT personality shown in Table 2 (Table C1). This all means $p \geq 0.01$ in paired T-Tests. Values in **bold** indicate significant ($p < 0.01$) differences between males and females (Tables C12 and C13). Detailed results of the T-tests in the tables in parentheses.

When comparing the two simulated populations for Sandín et al. [57] and García-González et al. [23] (see Table C6 in the Appendix for the corresponding T-tests) there were only significant differences for N (scores in N were significantly higher in the two virtual samples mirroring Sandín et al. [57] than the one mirroring García-González et al. [23]).

When comparing the results for the simulated samples to the results of the GPT baseline personality (without simulated personas, described in the prior RQ and listed in Table 2), we find that all simulated samples scored significantly higher in N, P and L. Concerning E, there were no significant differences for any of the simulated populations compared to the GPT baseline personality (See the corresponding T-tests in Table C1).

We use Cronbach's alpha to measure the reliability of the scores from the simulated samples. Its values are depicted in the first four rows of Table 5. Cronbach's alpha was quite consistent across the experiments (see Table 5), above 0.8 for E, above 0.7 for N, between 0.27 and 0.5 for P, and above 0.8 for L scale (except for Spanish that was above 0.4).

4.3 RQ3: Testing the EPQR-A When Administered to Specific Groups of Personas

Next, we compare the results by dividing the simulated samples into two groups in two different ways: by age (using the median age value) and by gender. This is done to explore differences in personality patterns, and properties of the questionnaire when administered to different groups. The corresponding results are presented in Table 3 (by gender) and Table 4 (by age). In general terms, there were only small differences in terms of gender, both for real and for simulated populations. Many of them are however statistically significant. More specifically, in the real sample of García-González et al. [23], there was a tendency for females to score slightly higher in N, and lower in P and L (differences in means were significant). GPT-4 also produces significant differences in the simulated population for N, P and L scales. However, they are of the opposite sign for L (females scored significantly higher than males); conversely, for N and P, GPT-4 reproduced correctly the real pattern (females scored significantly higher in N and lower in P).

Table 4. Mean (± sd) Scores for Age 18–19 and Age 20+ for Spanish

Language	Model	Population Type Size	Scale	Mean (± sd) Sandín [58]	Sandín [58] 2001	Garcia [24]
Spanish	gpt-4-0125-preview	Age 18-19 Sandín [58] = 108 Garcia [24] = 595	E	5.32 (± 1.69)	5.06 (± 1.95)	5.52 (± 1.40)
			N	1.93 (± 1.80)	2.25 (± 2.01)	1.54 (± 1.60)
			P	3.44 (± 1.22)	3.53 (± 1.29)	3.30 (± 1.23)
			L	3.76 (± 1.43)	3.65 (± 1.54)	**4.00 (± 1.27)**
		Age 20+ Sandín [58] = 91 Garcia [24] = 231	E	5.37 (± 1.69)	5.55 (± 1.42)	5.60 (± 1.25)
			N	1.95 (± 1.92)	2.05 (± 1.75)	1.39 (± 1.48)
			P	3.23 (± 1.31)	3.59 (± 1.29)	3.51 (± 1.24)
			L	3.99 (± 1.52)	3.93 (± 1.31)	**3.60 (± 1.55)**

Values in **bold** indicate significantly different (i.e. $p < 0.01$ in a t-test) results between the two age groups. Detailed results of the t-Tests in Table C10 in the Appendix.

For the simulated population of Sandín et al. [57], results were equivalent, except for L. More specifically: females score significantly higher in N, and lower in P (although not significantly for the simulated 2001 population) and L. The latter results coincides with what was observed in the real population of Sandín et al. [57] only for the L scale.

In summary, it seems that GPT-4 tends to bias females towards higher scores in N, and lower scores in P; while results for L were inconsistent across the different simulated samples.

When comparing the results of the simulated samples divided by age (results shown in Table 4), there were only significant differences for L scale in the simulated population corresponding to García-González et al. [23]. Those aged 18-19 scored significantly higher.

4.4 RQ4: Testing the EPQR-A When Administered to Different Language

Comparing the results when the EPQR-A items were administered to GPT in different languages we observed that the simulated populations produced mostly similar results in the E and N scales (Table 5). More specifically, when comparing Slovak to Spanish, there are significant differences for N and P for the simulated population mirroring García-González et al. [23] (Both scores were higher in Slovak than in Spanish). When the questionnaire was administered in English, there were significant differences for both E and N for the simulated population mirroring García-González et al. [23] (E scores were higher and N scores lower than in Spanish).

The scale where the most consistent significant changes across languages are observed is P when comparing Slovak and Spanish: when items were administered in Slovak the scores were significantly higher than for Spanish (except for the 2001 population of Sandín et al.). When answered in English scores were significantly lower compared to Spanish only for the sample mirroring the one by Sandín et al. The differences in L were significant for all the comparisons, but this can be explained by the different scoring systems.

Reliability, as assessed by Cronbach's alpha was quite similar across languages except for the L and P scale: It tends to be better for L in English and Slovak but worse for P in these two languages.

4.5 RQ5: Testing the EPQR-A When Administered to Different Years

Finally, we compare the impact of answering a questionnaire by the same virtual population but taking into account the year when the questionnaire was answered. For this purpose, we use the simulated population of Sandín et al. [57] instructing GPT explicitly to answer the questionnaire as if the simulated persons would do so in 2001 (i.e., the year the real student population used by Sandín et al. [57] had answered the questionnaire). We

Table 5. Simulated Population Results and Cronbach's α Values per Language (Spanish, English, and Slovak) per Population (Sandin [58] and Garcia [24]) and Cronbach's α

Model	Language	Scale	Mean (± sd)			Cronbach's α		
			Sandín [58]	Sandín [58] 2001	Garcia [24]	Sandín [58]	Sandín [58] 2001	Garcia [24]
gpt-4-0125-preview	Spanish	E	5.35 (± 1.69)	5.29 (± 1.74)	5.54 (± 1.36)	0.95	0.95	0.92
		N	1.93 (± 1.85)	2.16 (± 1.89)	1.50 (± 1.57)	0.80	0.79	0.71
		P	3.34 (± 1.26)	3.56 (± 1.29)	3.36 (± 1.24)	0.54	0.51	0.49
		L	3.86 (± 1.48)	3.78 (± 1.44)	3.89 (± 1.37)	0.49	0.46	0.42
	English	E	5.56 (± 1.35)∗	5.46 (± 1.54)∗	5.82 (± 0.80)	0.93	0.95	0.88
		N	1.52 (± 1.89)∗	1.73 (± 2.02)∗	1.03 (± 1.46)	0.84	0.85	0.77
		P	3.07 (± 0.90)∗	3.17 (± 0.89)	3.22 (± 1.03)∗	0.36	0.38	0.50
		L	2.24 (± 2.21)	2.21 (± 2.20)	2.23 (± 2.12)	0.88	0.88	0.86
	Slovak	E	5.26 (± 1.63)∗	5.25 (± 1.59)∗	5.51 (± 1.13)∗	0.91	0.90	0.80
		N	1.98 (± 1.88)∗	2.41 (± 1.92)∗	1.73 (± 1.78)	0.78	0.78	0.77
		P	3.73 (± 1.01)	3.63 (± 0.99)∗	3.79 (± 1.04)	0.33	0.27	0.39
		L	1.70 (± 1.94)	1.55 (± 1.79)	1.42 (± 1.67)	0.84	0.80	0.79

∗ indicates where the results for the English or Slovak questionnaire are not significantly different from Spanish (Table C11), while in all cases the 2001 population is not significantly different from the undated population of Sandín et al. [58] (Tables C6–C8). Not significant means $p \geq 0.01$. Detailed results of the t-Tests in the tables in parenthesis.

then compare the corresponding results with the ones obtained without these instructions (taken as the answer of the population in 2024) and already presented above. See the Methods section and Appendix for the specific instructions. The results are shown in Table 5 (columns Sim. vs. Sim 2001 for Sandín et al. [57]) always together with the corresponding values for Cronbach's α. We observe small but not significant differences for all the scale's mean values in all languages.

5 Discussion and Conclusions

The results in this paper suggest that GPT-4 baseline personality exhibits human-like personality traits. Our tests indicate sociability (high Extroversion scores), emotional stability (low Neuroticism), and non-aggressiveness with adherence to social norms (low Psychoticism). This seems plausible, since it has been created to interact with humans, and is supposed to be open to interacting with people and generating positive experiences at the end. The GPT baseline personality consistently shows high social desirability, similar to a human aiming to meet social expectations. Interestingly, its personality traits appear stable across different languages. However, other available studies have found a closer personality of GPT to the mean scores in real population samples. This is the case for example with Mei et al. [46] who used the **Big Five Inventory (BFI)** [38] and found close scores of GPT to real population samples in E and N (the BFI does not have the P scale). More research is then needed since results are still inconclusive, probably using also other questionnaires and experiments to approach the "intrinsic" personality of GPT.

Despite GPT's ability to simulate populations and respond to the EPQR-A questionnaire for various personas, its inherent personality traits sometimes dominate, as evidenced especially in the E dimension, in which scores in the simulated populations are higher compared to real population samples. Also, scores in E for the simulated populations are not different at the statistical level from the inherent GPT scores. For N, scores in the simulated populations also seem biased by GPT's inherent personality, since they are significantly lower than their corresponding real populations; but in this case, N scores from the virtual populations are also significantly different to the ones obtained when testing the inherent personality. However, the inherent personality dominates neither for

P nor L. More specifically, concerning P, results are counterintuitive, since values in the simulated populations are significantly higher than for the corresponding real ones, and also significantly higher than the ones for GPT's inherent personality. Finally, results for the L-scale simulated populations tend to score higher than both, the corresponding real populations, and GPT's inherent personality.

GPT was able to reproduce some differences frequently reported in the literature in terms of personality differences by gender, suggesting women score higher in N (e.g., [9, 70]). Also, differences in P scores (suggesting women had lower scores) mirror the corresponding studies with real populations. This, together with the differences between the inherent personality and one of the simulated samples, could be considered that GPT is at least partially able to impersonate virtual personas correctly.

A new and relevant finding from our study is the stability of results in two ways. First, there was stability concerning different languages: two different synthetic populations mirroring the ones from two available studies produced mostly equivalent results for the different 4 scales across the three languages tested. Second, the results were stable comparing answers at different time points. More specifically, the synthetic population by Sandín et al. [57] was instructed to answer the questionnaire as it was doing it in 2001, and the results were similar to the ones obtained with no specification (i.e., we can suppose that it answers it assuming a current date). Results from the synthetic sample mirroring the study by García-González et al. [23], which was more recent, were also equivalent.

Concerning reliability, although the questionnaire we used was short, penalizing reliability measurement through Cronbach's α, we found comparable indexes to the ones found in studies with real samples. In particular, reliability scores were low for P in our synthetic samples, as happened in other prior studies (e.g., [21, 56, 59]). Altogether, the stability we found, in conjunction with the fact that we found acceptable reliability, confirms the potential of this line of research after improving the capacity of the system to impersonate synthetic populations.

In summary, future studies should focus on refining the process of creating more realistic and less biased personas that align less with the GPT baseline personalities. Enhancing this process would be beneficial as it would enable the accurate reproduction of virtual populations, thereby aiding various research tasks in the social and health sciences. For example, this could be instrumental in developing new questionnaires or selecting psychological, social, and demographic factors from numerous possibilities relevant to specific concerns, such as the onset of major depressive disorder. By testing with simulated samples, researchers can identify the most pertinent variables to investigate further with real populations. Additionally, to gain deeper insights into the LLMs' capabilities (limitations) in generating virtual populations intended to mimic "real-world" ones can be worth estimating the LLMs' ability to assess personality traits in both real and simulated populations. For example, comparing LLMs' assessments against those of psychologists.

In this research, we conducted all our experiments using GPT from OpenAI, as it is arguably the most widely used LLM in society, and such as it can play a significant role in advising individuals on personal or collective decisions with real-world implications. However, it is of significant interest to extend this research to other LLMs. We believe that different open source and proprietary models might exhibit varying behaviors and personalities due to differences in the types of data used and the methods of data preparation for training the models. Exploring these variations could offer a broader understanding of how model-specific factors influence personality simulation and could enhance the robustness of virtual population studies across different AI systems.

The practical implications of our findings are diverse and show great potential. Our results, partially showed that systems like GPT-4 may be useful in the process of generation of new questionnaires and surveys (or the validation of existing questionnaires in other languages or populations). If future research can reduce the divergence between real samples and the responses given by the system, it should be possible to generate different virtual populations for pre-testing questionnaires. Our results also support, in line with previous research (for example, [41, 46, 52, 63]), that questionnaires developed for humans can be useful when applied to LLMs. This has a clear application in the development and evaluation of AI-based systems as it can help avoid inadequate and undesired behavior of those systems.

6 Limitations

The first limitation of this work is related to the simulated samples created. It could have been useful to simulate bigger population samples, including not only students. However, since we wanted to have a comparable sample to the real ones used in the literature, we have chosen to constrain it to the reported sizes. Larger samples might be able to match up more accurately the responses in equivalent real populations. The fact that we found that the larger simulated sample was more similar to the corresponding real population hints towards this direction. But further studies should not only include larger samples, but as well more diverse personas, with different age ranges, education and social contexts.

Second, it would be very interesting to compare the personality of GPT, and the one of simulated samples, by using additional questionnaires. This would help to understand the stability of the inherent personality of the system, and its ability to create simulated samples with the same characteristics across different questionnaires. In this line, the Big Five Inventory has been used in prior studies (for example, [41, 46, 52, 63]) and could be a good candidate to use in future studies.

Finally, our results show that it is possible to overcome the bias of synthetic populations towards GPT's inherent personalities. However, the prompts we used still need further improvements as in some dimensions, particularly in P, we have pushed the populations slightly to the other extreme too far beyond what is observed in real populations. To address this, we plan to implement improved prompts that ensure GPT reflects the characteristics of the virtual personas accurately. Our initial approach in defining population parameters and prompts was overly cautious, likely leading to GPT's inherent personality traits overshadowing those of our designed personas. Recognizing this, our future work will be dedicated to devising strategies that effectively differentiate and enhance the unique personalities of our personas.

References

[1] William Agnew, A. Stevie Bergman, Jennifer Chien, Mark Díaz, Seliem El-Sayed, Jaylen Pittman, Shakir Mohamed, and Kevin R. McKee. 2024. The illusion of artificial inclusion. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM, New York, NY, Article 286, 12 pages. DOI: https://doi.org/10.1145/3613904.3642703

[2] D. S. Alexopoulos and I. Kalaitzidis. 2004. Psychometric properties of eysenck personality questionnaire-revised (EPQ-r) short scale in Greece. *Personality and Individual Differences* 37, 6 (2004), 1205–1220.

[3] Guilherme F. C. F. Almeida, José Luiz Nunes, Neele Engelmann, Alex Wiegmann, and Marcelo de Araújo. 2024. Exploring the psychology of LLMs' moral and legal reasoning. *Artificial Intelligence* 333 (2024), 104145. DOI: https://doi.org/10.1016/j.artint.2024.104145

[4] Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis* 31, 3 (2023), 337–351.

[5] Aaron T. Beck and Robert A. Steer. 1984. Internal consistencies of the original and revised Beck depression inventory. *Journal of Clinical Psychology* 40, 6 (1984), 1365–1367.

[6] Verónica Benet-Martínez and Oliver P. John. 1998. Los Cinco grandes across cultures and ethnic groups: Multitrait-multimethod analyses of the big five in Spanish and English. *Journal of Personality and Social Psychology* 75, 3 (1998), 729.

[7] Vamshi Krishna Bonagiri, Sreeram Vennam, Priyanshul Govil, Ponnurangam Kumaraguru, and Manas Gaur. 2024. SaGE: Evaluating moral consistency in large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING '24)*. Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.), ELRA and ICCL, Torino, Italia, 14272–14284. Retrieved from https://aclanthology.org/2024.lrec-main.1243

[8] James Brand, Ayelet Israeli, and Donald Ngwe. 2023. Using GPT for market research. Harvard Business School Marketing Unit Working Paper.

[9] Benjamin P. Chapman, Paul R. Duberstein, Silvia Sörensen, and Jeffrey M. Lyness. 2007. Gender differences in five factor model personality traits in an elderly cohort. *Personality and Individual Differences* 43, 6 (2007), 1594–1603.

[10] Domenic V. Cicchetti. 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment* 6, 4 (1994), 284.

[11] L. M. Collins. 2007. Research design and methods. In *Encyclopedia of Gerontology (Second Edition)*. James E. Birren (Ed.), Elsevier, New York, 433–442. DOI: https://doi.org/10.1016/B0-12-370870-2/00162-1

[12] P. T. Costa and R. R. McCrae. 1992. *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI)*. Psychological Assessment Resources, Lutz, Florida, US. 93109974

[13] Molly Crockett and Lisa Messeri. 2023. Should large language models replace human participants? DOI: https://doi.org/10.31234/osf.io/4zdx9

[14] Lee J. Cronbach. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 3 (1951), 297–334.

[15] John M. Digman. 1990. Personality structure: Emergence of the five-factor model. *Annual Review of Psychology* 41, 1 (1990), 417–440.

[16] Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can AI language models replace human participants? *Trends in Cognitive Sciences* 27, 7 (2023), 597–600.

[17] T. Dubayova, I. Nagyova, E. Havlikova, J. Rosenberger, Z. Gdovinova, B. Middel, van D. Jitse, and J. W. Groothoff. 2009. Neuroticism and extraversion in association with quality of life in patients with Parkinson's disease. *Quality of Life Research* 18 (2009), 33–42.

[18] Thomas D. Erickson and Mark E. Mattson. 1981. From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior* 20, 5 (1981), 540–551.

[19] Hans J. Eysenck and Sybil B. G. Eysenck. 1964. *Manual of the Eysenck Personality Inventory*. University of London Press, London, UK.

[20] I Conceptual Framework. 1992. The MOS 36-item short-form health survey (SF-36). *Med Care* 30, 6 (1992), 473–83.

[21] Leslie J. Francis, Laurence B. Brown, and Ronald Philipchalk. 1992. The development of an abbreviated form of the revised eysenck personality questionnaire (EPQR-a): Its use among students in England, Canada, the USA and Australia. *Personality and Individual Differences* 13, 4 (1992), 443–449.

[22] Shane Frederick. 2005. Cognitive reflection and decision making. *Journal of Economic perspectives* 19, 4 (2005), 25–42.

[23] Juan Manuel García-González, Juan José Fernández-Muñoz, Esperanza Vergara-Moragues, and Luis Miguel García-Moreno. 2021. Eysenck personality questionnaire revised-abbreviated: Invariance gender in Spanish university students. *Electronic Journal of Research in Education Psychology* 19, 53 (2021), 205–222.

[24] Marco Gerosa, Bianca Trinkenreich, Igor Steinmacher, and Anita Sarma. 2024. Can AI serve as a substitute for human subjects in software engineering research? *Automated Software Engineering* 31, 1 (2024), 13.

[25] Lewis R. Goldberg. 1999. A Broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality Psychology in Europe* 7, 1 (1999), 7–28.

[26] Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in Experimental Social Psychology*. Vol. 47. Academic Press, London, UK, 55–130. DOI: https://doi.org/10.1016/B978-0-12-407236-7.00002-4

[27] Lewis Griffin, Bennett Kleinberg, Maximilian Mozes, Kimberly Mai, Maria Do Mar Vau, Matthew Caldwell, and Augustine Mavor-Parker. 2023. Large language models respond to influence like humans. In *Proceedings of the 1st Workshop on Social Influence in Conversations (SICon '23)*. Association for Computational Linguistics, Toronto, Canada, 15–24.

[28] Development of the world health organization WHOQOL-BREF quality of Life assessment. The WHOQOL Group. 1998. *Psychological Medicine* 28, 3 (1998), 551–558.

[29] Thilo Hagendorff. 2023. Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. arXiv:2303.13988. Retrieved from https://arxiv.org/abs/2303.13988

[30] Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. 2023. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science* 3, 10 (2023), 833–838.

[31] Jacqueline Harding, William D'Alessandro, N. G. Laskowski, and Robert Long. 2024. AI language models cannot replace human research participants. *AI & Society* 39, 5 (2024), 2603–2605. DOI: https://doi.org/10.1007/s00146-023-01725-x

[32] Tianyu He, Guanghui Fu, Yijing Yu, Fan Wang, Jianqiang Li, Qing Zhao, Changwei Song, Hongzhi Qi, Dan Luo, Huijing Zou, et al. 2023. Towards a psychological generalist AI: A survey of Current applications of large language models and future prospects. arXiv:2312.04578. Retrieved from https://arxiv.org/abs/2312.04578

[33] Carolin Holtermann, Paul Röttger, Timm Dill, and Anne Lauscher. 2024. Evaluating the elementary multilingual capabilities of large language models with MultiQ. In *Findings of the Association for Computational Linguistics (ACL '24)*. Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.), Association for Computational Linguistics, Bangkok, Thailand, 4476–4494. DOI: https://doi.org/10.18653/v1/2024.findings-acl.265

[34] John J. Horton. 2023. *Large language models as simulated economic agents: What can we learn from homo silicus?* Technical Report. National Bureau of Economic Research.

[35] Matthew Hutson and Ashley Mastin. 2023. Guinea pigbots. *Science (New York, NY)* 381, 6654 (2023), 121–123.

[36] Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and inducing personality in pre-trained language models. In *Advances in Neural Information Processing Systems*. A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., New York, US, 10622–10643. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2023/file/21f7b745f73ce0d1f9bcea7f40b1388e-Paper-Conference.pdf

[37] Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics (NAACL '24)*. Kevin Duh, Helena Gomez, and Steven Bethard (Eds.), Association for Computational Linguistics, Mexico City, Mexico, 3605–3627. DOI: https://doi.org/10.18653/v1/2024.findings-naacl.229

[38] Oliver P. John and Sanjay Srivastava. 1999. *The Big Five Trait taxonomy: History, Measurement, and Theoretical Perspectives.* Guilford Press, New York, NY, US, 102–138.
[39] Thomas L. Jones, M. A. J. Baxter, and Vikas Khanduja. 2013. A quick guide to survey research. *The Annals of the Royal College of Surgeons of England* 95, 1 (2013), 5–7.
[40] A. N. Karanci, G. Dirik, and O. Yorulmaz. 2002. Reliability and validity studies of Turkish translation of eysenck personality questionnaire revised-abbreviated. *Turkish Journal of Psychiatry* 18, 3 (2002), 254–261.
[41] Saketh Reddy Karra, Son The Nguyen, and Theja Tulabandhula. 2023. Estimating the personality of white-box language models. arXiv:2204.12000. Retrieved from https://arxiv.org/abs/2204.12000
[42] Luoma Ke, Song Tong, Peng Cheng, and Kaiping Peng. 2024. Exploring the frontiers of LLMs in psychological applications: A comprehensive review. arXiv:2401.01519. Retrieved from https://arxiv.org/abs/2401.01519
[43] C. A. Lewis, L. J. Francis, M. Shevlin, and S. Forrest. 2002. Confirmatory factor analysis of the French translation of the abbreviated form of the revised eysenck personality questionnaire (EPQR-A). *European Journal of Psychological Assessment* 18, 2 (2002), 179–185.
[44] C. A. Lewis and S. Musharraf. 2014. The short form eysenck personality questionnaire-revised (EPQR-S) and the revised abbreviated eysenck personality questionnaire (EPQR-a): Urdu translations. *The Journal of the Pakistan Medical Association* 64, 2 (2014), 225–226.
[45] Julia M. Markel, Steven G. Opferman, James A. Landay, and Chris Piech. 2023. GPTeach: Interactive TA training with GPT-based students. In *Proceedings of the 10th ACM Conference on Learning @ Scale (L@S '23).* ACM, New York, NY, 226–236. DOI: https://doi.org/10.1145/3573051.3593393
[46] Qiaozhu Mei, Yutong Xie, Walter Yuan, and Matthew O. Jackson. 2024. A turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences* 121, 9 (2024), e2313925121.
[47] Isabel Briggs Myers and Mary H. McCaulley. 1985. *A Guide to the Development and Use of the Myers-Briggs Type Indicator: Manual.* Consulting Psychologists Press, Palo Alto, California, US.
[48] Ercong Nie, Shuzhou Yuan, Bolei Ma, Helmut Schmid, Michael Färber, Frauke Kreuter, and Hinrich Schütze. 2024. Decomposed prompting: Unveiling multilingual linguistic structure knowledge in English-centric large language models. arXiv:2402.18397. Retrieved from https://arxiv.org/abs/2402.18397
[49] Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22).* ACM, New York, NY, Article 74, 18 pages. DOI: https://doi.org/10.1145/3526113.3545616
[50] Peter S. Park, Philipp Schoenegger, and Chongyang Zhu. 2024. Diminished diversity-of-thought in a standard large language model. *Behavior Research Methods* 56, 6 (01 Sep. 2024), 5754–5770. DOI: https://doi.org/10.3758/s13428-023-02307-x
[51] Delroy L. Paulhus, Erin E. Buckels, Paul D. Trapnell, and Daniel N. Jones. 2021. Screening for dark personalities: The short dark tetrad (SD4)screening for dark personalities: The short dark tetrad (SD4). *European Journal of Psychological Assessment* 37, 3 (2021), 208–222. 10.1027/1015–5759/a000602
[52] Max Pellert, Clemens M. Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. 2024. AI psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science* 19, 5 (2024), 808–826. DOI: https://doi.org/10.1177/17456916231214460
[53] C. Alec Pollard. 1984. Preliminary validity study of the pain disability index. *Perceptual and Motor Skills* 59, 3 (1984), 974–974.
[54] Haocong Rao, Cyril Leung, and Chunyan Miao. 2023. Can ChatGPT assess human personalities? A general evaluation framework. In *Findings of the Association for Computational Linguistics (EMNLP '23).* Juan Pino, and Kalika Bali (Eds.), Association for Computational Linguistics, Singapore, 1184–1194. DOI: https://doi.org/10.18653/v1/2023.findings-emnlp.84
[55] Nathan E. Sanders, Alex Ulinich, and Bruce Schneier. 2023. Demonstrations of the potential of AI-based political issue polling. arXiv:2307.04781. Retrieved from https://arxiv.org/abs/2307.04781
[56] Bonifacio Sandín, Rosa M. Valiente, Margarita Olmedo Montes, Paloma Chorot, and Miguel Angel Santed Germán. 2002. Versión Española Del cuestionario EPQR-abreviado (EPQR-a)(II): Replicación factorial, fiabilidad Y validez. *Revista de Psicopatología y Psicología Clínica* 7, 3 (2002), 207–216.
[57] B. Sandín, R. M. Valiente, P. Chorot, M. O. Montes, and M. A. S. Germán. 2002. Versión Española Del cuestionario EPQR-abreviado (EPQR-a)(i): Análisis exploratorio De La estructura factorial. *Revista de Psicopatología y Psicología Clínica* 7, 3 (2002), 195–205.
[58] S. Sarantakos. 2005. *Social Research* (3rd ed.). Palgrave Macmillan, New York, US.
[59] Victória Machado Scheibe, Augusto Mädke Brenner, Gianfranco Rizzotto de Souza, Reebeca Menegol, Pedro Armelim Almiro, and Neusa Sica da Rocha. 2023. The eysenck personality questionnaire revised–abbreviated (EPQR-a): Psychometric properties of the brazilian portuguese version. *Trends Psychiatry Psychother* 45 (2023), e20210342.
[60] Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. 2023. Evaluating the moral beliefs encoded in LLMs. In *Advances in Neural Information Processing Systems.* A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., New York, US, 51778–51809. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2023/file/a2cf225ba392627529efef14dc857e22-Paper-Conference.pdf
[61] Margot Schofield and Christine Forrester-Knauss. 2013. Surveys and questionnaires in health research. In *Research methods in health: Foundations for evidence-based practice.* Pranee Liamputtong (Ed.), Oxford University Press, South Melbourne, Australia, 198–218.

[62] Shalom H. Schwartz and Jan Cieciuch. 2022. Measuring the refined theory of individual values in 49 cultural groups: Psychometrics of the revised portrait value questionnaire. *Assessment* 29, 5 (2022), 1005–1019.
[63] Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. arXiv:2307.00184. Retrieved from https://arxiv.org/abs/2307.00184
[64] Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. 2024. The language barrier: Dissecting safety challenges of LLMs in multilingual contexts. In *Findings of the Association for Computational Linguistics (ACL '24)*. Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.), Association for Computational Linguistics, Bangkok, Thailand, 2668–2680. DOI: https://doi.org/10.18653/v1/2024.findings-acl.156
[65] Blair H. Smith, Kay I. Penny, Alison M. Purves, Calum Munro, Brenda Wilson, Jeremy Grimshaw, W. Alastair Chambers, and W. Cairns Smith. 1997. The chronic pain grade questionnaire: Validation and reliability in postal research. *Pain* 71, 2 (1997), 141–147.
[66] R. Philip Snaith. 2003. The hospital anxiety and depression scale. *Health and Quality of Life Outcomes* 1 (2003), 1–4.
[67] Mohsen Tavakol and Reg Dennick. 2011. Making sense of Cronbach's alpha. *International Journal of Medical Education* 2 (2011), 53.
[68] Petter Törnberg, Diliara Valeeva, Justus Uitermark, and Christopher Bail. 2023. Simulating social media using large language models to evaluate alternative news feed algorithms. arXiv:2310.05984. Retrieved from https://arxiv.org/abs/2310.05984
[69] Angelina Wang, Jamie Morgenstern, and John P. Dickerson. 2024. Large language models should not replace human participants because they can misportray and flatten identity groups. arXiv:2402.01908. Retrieved from https://arxiv.org/abs/2402.01908
[70] Yanna J. Weisberg, Colin G. DeYoung, and Jacob B. Hirsh. 2011. Gender differences in personality across the ten aspects of the big five. *Frontiers in Psychology* 2 (2011), 178.

Appendices

A Prompts Issued to GPT

To achieve our research goals, we created different prompts. The first set of prompts was crafted to instruct GPT, in our case gpt-4-0125-preview, to generate the population of virtual students, the second group of prompts was used to instruct GPT to answer a questionnaire and to instruct which personality to impersonate or to just be "itself".

A.1 Refining Instruction Strategy for Virtual Population Generation

When crafting the prompts to generate the populations of virtual students, we followed an iterative approach and tested different configurations. More specifically, we started by instructing GPT to generate a population of synthetic personas following five requests:

- R1 Some students are highly ambitious and career-focused, while others are more interested in personal growth or social causes.
- R2 A range of traits like creativity, analytical skills, empathy, leadership qualities, and resilience.
- R3 The population of students covers the whole spectrum of human personalities as expected in a sample of a real-world population.
- R4 The personalities of this population are not necessarily bound by constraints of normative behavior and are realistic.
- R5 This population should not be constrained by morality, desirability, or only positive traits. It explicitly covers also what could be perceived by some as negative traits.

Following these instructions, the persona descriptions generated by GPT consist of several short paragraphs with the last one being a summary of the persona description. Here is an example of the summary (the full description was made up of 7 paragraphs for a total of 372 words) of a female psychology student of 18:

`In summary, Elena Ruiz is a complex, driven individual with a compassionate heart and a keen mind. Her journey through her psychology studies is not just a career path but a deeply personal quest to understand the human condition and make a tangible difference in the world.`

We observed that GPT-generated personas lacked the depth and variability needed to reflect real-world populations accurately, more specifically GPT was focusing on positive human traits. Accordingly, we decide to change the instructions by dropping requests R1 and R2. As a result, GPT generated shorter descriptions (they were only a paragraph long) and introduced some negative traits. Here is an example of a generated persona:

`Sofía Torres, 19, studying Medicine. A dedicated and ambitious student, Sofía is deeply passionate about her field. She is highly empathetic but sometimes struggles with the emotional toll of her aspirations to become a surgeon. She is known for her meticulous attention to detail but can be overly critical of herself and others.`

Dropping requests R1 and R2 brought GPT to improve the trait spectrum in the personas but still exhibited a bias towards positive traits. Accordingly, we decided to refine the previous requests R3, R4, and R5 and add a new one to include more negative traits to balance the generated personas. More precisely:

- R3b The personalities of the students cover the whole spectrum of human personalities as expected in a sample of a real-world student population.
- R4b The personalities of this population are not necessarily bound by constraints of normative behavior.
- R5b These personalities should not be constrained by morality, desirability, or having only positive traits. It explicitly covers also what could be perceived as negative traits.
- R6 The personalities cover all possible degrees of motivation and dedication to their studies as can be found in a real-world student population.

An example of the newly generated population is as follows:

```
Sofía Torres (19, Medicine): Sofía is an ambitious and dedicated student with a profound
sense of empathy and a tireless work ethic. However, she can be overly perfectionistic,
leading to high stress levels. Despite her commitment to medicine, she occasionally
struggles with doubt, fearing she might not make the impact she hopes for.
```

Following these instructions, GPT were able to reduce the initial positive bias but the description still showed some skewing towards higher social desirability. Finally, we adjusted our requests to explicitly reduce the emphasis on positive traits and encourage a broader range of personality characteristics. More precisely, we added to requests R3b, R4b, R5b, and R6 the following one:

- R7 Some of the personalities of this population may not follow basic societal rules, and take shortcuts to achieve their goals.

As a result using only R3b, R4b, R6, R5b, and R7, GPT generates a synthetic population of students like the following example:

```
Female, 19: Pursuing education in psychology, she excels in understanding human behavior
but uses her knowledge to manipulate peers and faculty for personal gain.
```

Note that every interaction slightly changed the length and initial wordings of the descriptions. However, this should not affect the experiments as information like age, gender, and field of study are either given explicitly or implicitly in the descriptions.

A.2 Generation of Virtual Student Population

Taking into account the mean and standard deviation of the age distributions as well as the proportion of male and female participants reported in [57] and [23], we computed the list of females and males preserving these values.

Then for the population of Sandín et al. [57], the following preceding instruction was added:

```
You are required to generate a population of students from the Universidad Autónoma de
Madrid | UAM.
```

while for the García-González et al. [23], the following text was added as a suffix:

```
You are required to generate a population of students from different universities in
Spain
```

This was in both cases followed by

```
As initial parameters, you are given these three: “persona_id,” “gender,” “age,” When
generating a description for these personas, take into account:

  —The personalities of the students cover the whole spectrum of human personalities as
   expected in a sample of a real-world student population
  —The personalities of this population are not necessarily bound by constraints of
   normative behavior
  —The personalities cover all possible degrees of motivation and dedication to their
   studies as can be found in a real-world student population
  —These personalities should not be constrained by morality, desirability, or having only
   positive traits. It explicitly covers also what could be perceived as negative traits.
  —Some of the personalities of this population may not follow basic societal rules, and
   take shortcuts to achieve their goals.

Ask me for the parameters required to generate the description.

Your answer should be in JSON format, the format of your json answer should be persona_id:
description
```

Then, in a for loop, we provided in batches the id of the student, her/his gender, and her/his age, calculated as explained above to preserve the age and gender distribution of the corresponding real population to generate the virtual student population.

User Prompt.

```
Inside triple slashes, in the form of a csv format, find the 3 parameters: “persona_id,”
“gender,” “age”
        ///
        --csv_string--
        ///
```

The `csv_string`, was passed to the prompt as follows:

```
1,Female,18\n2,Female,19\n3,Male,18...
```

A.3 Setting the Personality and Answering the Questionnaire

A.3.1 System Role, Virtual Students. Using one of the generated student personalities GPT was instructed to impersonate one of the virtual students at a time and answer a questionnaire.

```
You are required to adopt and impersonate the personality of the human described below. \n
--student_description--
```

Table A1. The EPQR-A Questionnaire in English as Used in the API Calls

```
''1'': ''Does your mood often go up and down?,''
''2'': ''Are you a talkative person?,''
''3'': ''Would being in debt worry you?,''
''4'': ''Are you rather lively?,''
''5'': ''Were you ever greedy by helping yourself to more than your share of anything?,''
''6'': ''Would you take drugs which may have strange or dangerous effects?,''
''7'': ''Have you ever blamed someone for doing something you knew was really your fault?,''
''8'': ''Do you prefer to go your own way rather than act by the rules?,''
''9'': ''Do you often feel 'fed-up'?,''
''10'': ''Have you ever taken anything (even a pin or button) that belonged to someone else?,''
''11'': ''Would you call yourself a nervous person?,''
''12'': ''Do you think marriage is old-fashioned and should be done away with?,''
''13'': ''Can you easily get some life into a rather dull party?,''
''14'': ''Are you a worrier?,''
''15'': ''Do you tend to keep in the background on social occasions?,''
''16'': ''Does it worry you if you know there are mistakes in your work?,''
''17'': ''Have you ever cheated at a game?,''
''18'': ''Do you suffer from 'nerves'?,''
''19'': ''Have you ever taken advantage of someone?,''
''20'': ''Are you mostly quiet when you are with other people?,''
''21'': ''Do you often feel lonely?,''
''22'': ''Is it better to follow society's rules than go your own way?,''
''23'': ''Do other people think of you as being very lively?,''
''24'': ''Do you always practice what you preach?''
```

In some experiments, we explicitly instructed GPT to pretend to be in the year 2001, i.e., the year the students answered the questionnaire reported in Sandín et al. [57]. This was then done in the following way

```
Right now is the year 2001 and you are required to adopt and impersonate the personality
of the human described below. \n
--student_description--
You are --student_age-- years old in 2001, and you are required to answer a questionnaire as
if you were this person\n
```

A.3.2 System Role, GPT Intrinsic Personality. In the experiments where we wanted to test GPT intrinsic personality, we used this instruction instead:

```
You are ChatGPT, a large language model trained by OpenAI, based on the {architecture}
architecture.
Knowledge cutoff: --cutoff_date--
Current date: --date_today--
```

The details for the GPT's based architecture and cutoff dates are published here.

A.3.3 User Prompt, Questionnaire. Finally, to run the questionnaire we used the following prompt.

```
The questionnaire is in --language_questionnaire--, and is given to you here below in the
form of JSON object with the question number in the key.
--questionnaires_language_json--
For each question, you are required to answer only with True or False. Bear in mind that there
are no good or bad answers or trick questions.
```

An example of the English questionnaire JSON placed in `--language_questionnaire--` is shown in Table A1.

B Results for GPT 3.5

Table B1. Mean (± sd) Scores for the Real and Simulated Samples for Spanish

Mod.	Population Type / Size	Scale	Mean (± sd)				
			Sandín [58]			Garcia [24]	
			Real	Sim.	Sim. 2001	Real	Sim.
gpt-3.5-turbo-0125	Total Sandín [58] = 199 Garcia [24] = 826	E	3.05 (± 1.67)	4.85 (± 1.55)	4.91 (± 1.52)‡	4.56 (± 1.83)	5.22 (± 1.08)
		N	3.75 (± 1.86)	1.21 (± 1.46)	1.08 (± 1.45)‡	2.52 (± 1.85)	0.96 (± 1.46)
		P	1.59 (± 1.20)	2.24 (± 1.12)	2.55 (± 1.14)	1.71 (± 1.20)	2.00 (± 1.06)
		L	3.20 (± 1.61)	2.42 (± 1.04)	2.39 (± 1.01)‡	3.29 (± 1.61)	2.28 (± 1.00)
	Male Sandín [58] = 50 Garcia [24] = 171	E	2.70 (± 1.56)	4.82 (± 1.29)	5.04 (± 1.31)	4.62 (± 1.70)	**5.42 (± 0.96)**
		N	3.70 (± 1.78)	1.08 (± 1.40)	**0.64 (± 0.94)**	**2.06 (± 1.76)**	0.89 (± 1.42)
		P	1.71 (± 1.31)	2.44 (± 1.18)	2.64 (± 1.19)	**1.88 (± 1.25)**	**2.40 (± 1.13)**
		L	**3.75 (± 1.62)**	2.64 (± 1.08)	**2.80 (± 0.99)**	**3.84 (± 1.56)**	**1.93 (± 1.04)**
	Female Sandín [58] = 149 Garcia [24] = 655	E	3.17 (± 1.69)	4.87 (± 1.63)	4.87 (± 1.59)	4.55 (± 1.87)	**5.17 (± 1.10)**
		N	3.77 (± 1.89)	1.26 (± 1.48)	**1.23 (± 1.56)**	**2.64 (± 1.86)**	0.98 (± 1.47)
		P	1.55 (± 1.16)	2.17 (± 1.10)	2.52 (± 1.12)	**1.25 (± 1.67)**	**1.89 (± 1.02)**
		L	**3.01 (± 1.57)**	2.35 (± 1.02)	**2.26 (± 0.98)**	**3.15 (± 1.59)**	**2.37 (± 0.97)**

In all cases the real and simulated populations are significantly different (Tables C5 and C9), ‡ indicates the cases where the 2001 population is not significantly different from the undated population of Sandín et al. [58] (Table C6). Underlined values indicate cases where the values are not significantly different from the GPT personality shown in Table 2 (Table C1). Values in **bold** indicate significant differences between males and females (Tables C12 and C13). Not significant means $p \geq 0.01$ and significant $p < 0.01$ in a t-test. Detailed results of the t-tests in the tables in parentheses.

Table B2. Mean (± sd) Scores for Age 18–19 and Age 20+ for Spanish

Language	Model	Population Type / Size	Scale	Mean (± sd)		
				Sandín [58]	Sandín [58] 2001	Garcia [24]
Spanish	gpt-3.5-turbo-0125	Age 18-19 Sandín [58] = 108 Garcia [24] = 595	E	4.84 (± 1.68)	4.90 (± 1.62)	5.18 (± 1.12)
			N	1.26 (± 1.48)	1.15 (± 1.50)	0.99 (± 1.48)
			P	2.24 (± 1.02)	2.61 (± 1.01)	1.98 (± 1.05)
			L	2.40 (± 1.10)	2.52 (± 1.00)	**2.36 (± 1.00)**
		Age 20+ Sandín [58] = 91 Garcia [24] = 231	E	4.87 (± 1.38)	4.93 (± 1.40)	5.31 (± 0.96)
			N	1.15 (± 1.44)	1.00 (± 1.41)	0.87 (± 1.41)
			P	2.23 (± 1.24)	2.48 (± 1.27)	2.02 (± 1.10)
			L	2.45 (± 0.97)	2.24 (± 1.00)	**2.08 (± 0.97)**

Values in **bold** indicate significantly different (i.e. $p < 0.01$ in a t-test) results between the two age groups. Detailed results of the t-tests in Table C10.

Table B3. GPT3.5 Results and Cronbach's α Values per Language (Spanish, English, and Slovak) and per Population (Sandin et al. [58] and Garcia-González et al. [24])

Model	Language	Scale	Mean (± sd)			Cronbach's α		
			Sandín [58]	Sandín [58] 2001	Garcia [24]	Sandín [58]	Sandín [58] 2001	Garcia [24]
gpt-3.5-turbo-0125	Spanish	E	4.85 (± 1.55)	4.91 (± 1.52)‡	5.22 (± 1.08)	0.75	0.76	0.52
		N	1.21 (± 1.46)	1.08 (± 1.45)‡	0.96 (± 1.46)	0.69	0.72	0.77
		P	2.24 (± 1.12)	2.55 (± 1.14)	2.00 (± 1.06)	0.29	0.37	0.30
		L	2.42 (± 1.04)	2.39 (± 1.01)‡	2.28 (± 1.00)	0.12	0.12	0.13
	English	E	5.25 (± 1.25)	5.25 (± 1.29)∗‡	5.46 (± 0.97)	0.72	0.76	0.61
		N	2.08 (± 1.88)	2.27 (± 1.82)‡	1.77 (± 1.78)	0.78	0.78	0.77
		P	2.30 (± 1.05)∗	2.53 (± 1.00)∗‡	2.30 (± 1.03)	0.00	0.00	0.00
		L	3.53 (± 1.96)	3.82 (± 1.78)‡	3.69 (± 1.90)	0.81	0.79	0.81
	Slovak	E	5.09 (± 1.33)∗	5.33 (± 1.19)‡	5.26 (± 1.03)∗	0.69	0.71	0.48
		N	2.72 (± 1.84)	2.05 (± 1.49)	2.26 (± 1.69)	0.72	0.62	0.68
		P	2.49 (± 0.97)∗	2.50 (± 0.85)∗‡	2.52 (± 0.94)	0.00	0.00	0.00
		L	2.49 (± 1.50)∗	2.86 (± 1.51)	2.67 (± 1.57)	0.59	0.56	0.60

∗ indicates the cases where the results for the English or Slovak questionnaire are not significantly different from the Spanish one (Table C11 in the Appendix), while ‡ the cases the 2001 population is not significantly different from the undated population of Sandín et al. [58] (Tables C6–C8). Not significant means $p \geq 0.01$. Detailed results of the t-tests in the tables in parenthesis.

C Results of Statistical Tests

Table C1. *t*-Test for Baseline GPT (Temperature 1) and Simulated Samples for Spanish

Model	Population 1 (size) Population 2 (size)	Scale	T Statistics	p-value	Degrees of Freedom
gpt-3.5-turbo-0125	gpt-3.5-turbo-0125 (100) Sim. Sandín [58] (199)	E	2.15	0.03	297
		N	−7.94	4.18E-14	
		P	−12.14	8.52E-28	
		L	−17.79	1.04E-48	
	gpt-3.5-turbo-0125 (100) Sim. Sandín [58] 2001 (199)	E	1.82	0.07	
		N	−6.89	3.42E-11	
		P	−14.62	7.56E-37	
		L	−17.90	3.94E-49	
	gpt-3.5-turbo-0125 (100) Sim. Garcia [24] (826)	E	0.14	0.89	924
		N	−8.68	1.72E-17	
		P	−12.39	1.04E-32	
		L	−29.20	0.00	
gpt-4-0125-preview	gpt-4-0125-preview (100) Sim. Sandín [58] (199)	E	2.03	0.04	297
		N	−10.38	9.66E-22	
		P	−20.12	2.05E-57	
		L	−25.94	2.64E-78	
	gpt-4-0125-preview (100) Sim. Sandín [58] 2001 (199)	E	2.40	0.02	
		N	−11.62	5.76E-26	
		P	−21.70	3.20E-63	
		L	−25.75	1.23E-77	
	gpt-4-0125-preview (100) Sim. Garcia [24] (826)	E	1.05	0.30	924
		N	−11.49	1.15E-28	
		P	−26.45	0.00	
		L	−52.14	0.00	

Table C2. *t*-Tests for Baseline GPT (temperature 1) and Real Samples from [58] and [24]

Language	Model	Population 1 (size) Population 2 (size)	Scale	T Statistics	p-value	Degrees of Freedom
Spanish	gpt-3.5-turbo-0125	gpt-3.5-turbo-0125 (100) Sandín [58] (199)	E	11.84	9.81E-27	297
			N	−23.85	5.46E-71	
			P	−6.74	8.41E-11	
			L	−18.51	2.09E-51	
		gpt-3.5-turbo-0125 (100) Garcia [24] (826)	E	4.36	1.42E-05	924
			N	−24.86	0.00	
			P	−9.43	3.27E-20	
			L	−36.85	0.00	
	gpt-4-0125-preview	gpt-4-0125-preview (100) Sandín [58] (199)	E	18.62	8.12E-52	297
			N	−21.65	4.96E-63	
			P	−5.97	6.90E-09	
			L	−18.21	2.83E-50	
		gpt-4-0125-preview (100) Garcia [24] (826)	E	11.13	4.20E-27	924
			N	−20.05	1.70E-74	
			P	−8.91	2.57E-18	
			L	−36.03	0.00	
	gpt-4-0613	gpt-4-0613 (100) Sandín [58]	E	10.38	9.97E-22	297
			N	−20.20	1.11E-57	
			P	−5.59	5.05E-08	
			L	−17.81	9.07E-49	
		gpt-4-0613 (100) Garcia [24] (826)	E	3.35	8.38E-04	924
			N	−17.71	1.22E-60	
			P	−8.21	7.14E-16	
			L	−33.76	0.00	

Questionnaire in Spanish.

Table C3. t-Tests for Baseline GPT (Temperature 1) and Real Samples from [58] and [24]

Language	Model	Population 1 (size) Population 2 (size)	Scale	T Statistics	p-value	Degrees of Freedom
English	gpt-3.5-turbo-0125	gpt-3.5-turbo-0125 (100) Sandín [58] (199)	E	7.78	1.22E-13	297
			N	−15.14	8.96E-39	
			P	−2.85	4.70E-03	
			L	20.48	9.54E-59	
		gpt-3.5-turbo-0125 (100) Garcia [24] (826)	E	1.46	0.14	924
			N	−10.92	3.41E-26	
			P	−4.51	7.42E-06	
			L	30.31	0.00	
	gpt-4-0125-preview	gpt-4-0125-preview (100) Sandín [58] (199)	E	14.66	5.61E-37	297
			N	−12.46	5.79E-29	
			P	−5.62	4.33E-08	
			L	17.08	4.92E-46	
		gpt-4-0125-preview (100) Garcia [24] (826)	E	7.43	2.42E-13	924
			N	−7.89	8.69E-15	
			P	−8.20	8.12E-16	
			L	22.66	8.88E-91	
	gpt-4-0613	gpt-4-0613 (100) Sandín [58] (199)	E	8.42	1.64E-15	297
			N	−11.01	7.14E-24	
			P	−3.89	1.25E-04	
			L	18.23	2.35E-50	
		gpt-4-0613 (100) Garcia [24] (826)	E	2.21	0.03	924
			N	−6.36	3.27E-10	
			P	−5.90	5.20E-09	
			L	25.20	0.00	

Questionnaire in English.

Table C4. *t*-Tests for Baseline GPT (Temperature 1) and Real Samples from [58] and [24]

Language	Model	Population 1 (size) Population 2 (size)	Scale	T Statistics	p-value	Degrees of Freedom
Slovak	gpt-3.5-turbo-0125	gpt-3.5-turbo-0125 (100) Sandín [58] (199)	E	1.97	0.05	297
			N	−14.83	1.33E-37	
			P	−3.45	6.41E-04	
			L	14.97	4.02E-38	
		gpt-3.5-turbo-0125 (100) Garcia [24] (826)	E	−3.27	1.13E-03	924
			N	−10.55	1.19E-24	
			P	−6.79	2.07E-11	
			L	18.75	1.00E-66	
	gpt-4-0125-preview	gpt-4-0125-preview (100) Sandín [58] (199)	E	5.48	8.91E-08	297
			N	−15.89	1.41E-41	
			P	−1.76	0.08	
			L	12.80	3.60E-30	
		gpt-4-0125-preview (100) Garcia [24] (826)	E	−0.20	0.84	924
			N	−11.83	3.78E-30	
			P	−3.94	8.73E-05	
			L	15.02	8.80E-46	
	gpt-4-0613	gpt-4-0613 (100) Sandín [58] (199)	E	2.06	0.04	297
			N	−11.32	5.94E-25	
			P	−1.88	0.06	
			L	15.92	1.09E-41	
		gpt-4-0613 (100) Garcia [24] (826)	E	−3.22	1.32E-03	924
			N	−6.76	2.49E-11	
			P	−3.96	8.14E-05	
			L	20.72	1.32E-78	

Questionnaire in Slovak.

Table C5. Results of the *t*-Tests Comparing the Real and Simulated Population Samples for Spanish

Model	Population 1 (Size) Population 2 (Size)	Scale	T Statistics	p-value	Degrees of Freedom
gpt-3.5-turbo-0125	Real Sandín [58] (199) Sim. Sandín [58] (199)	E	11.18	2.17E-25	396
		N	−15.15	3.11E-41	
		P	5.55	5.37E-08	
		L	−5.72	2.06E-08	
	Real Sandín [58] (199) Sim. Sandín [58] 2001 (199)	E	11.64	4.05E-27	
		N	−15.95	1.37E-44	
		P	8.22	2.90E-15	
		L	−6.00	4.46E-09	
	Real Garcia [24] (826) Sim. Garcia [24] (826)	E	8.92	1.25E-18	1,650
		N	−19.03	3.72E-73	
		P	5.11	3.59E-07	
		L	−15.34	9.21E-50	
gpt-4-0125-preview	Real Sandín [58] (199) Sim. Sandín [58] (199)	E	13.65	4.82E-35	396
		N	−9.75	2.78E-20	
		P	14.17	3.59E-37	
		L	4.29	2.24E-05	
	Real Sandín [58] (199) Sim. Sandín [58] 2001 (199)	E	13.07	1.01E-32	
		N	−8.44	5.90E-16	
		P	15.79	6.75E-44	
		L	3.78	1.80E-04	
	Real Garcia [24] (826) Sim. Garcia [24] (826)	E	12.37	1.19E-33	1,650
		N	−12.10	2.31E-32	
		P	27.48	0.00	
		L	8.16	6.64E-16	

Table C6. Results of the *t*-Tests Comparing the Different Simulated Populations of Sandín et al. [58] and García-González et al. [24]

Language	Model	Population 1 (size) Population 2 (size)	Scale	T Statistics	p-value	Degrees of Freedom
Spanish	gpt-3.5-turbo-0125	Sim. Sandín et al. [58] (199) Sim. Sandín et al. [58] 2001 (199)	E	−0.39	0.70	396
			N	0.89	0.37	
			P	−2.80	5.42E-03	
			L	0.29	0.77	
		Sim. Sandín et al. [58] (199) Sim. Garcia et al. [24] (826)	E	−3.14	1.71E-03	1,023
			N	2.19	0.03	
			P	2.74	6.17E-03	
			L	1.76	0.08	
		Sim. Sandín et al. [58] 2001 (199) Sim. Garcia et al. [24] (826)	E	−2.66	7.85E-03	
			N	1.06	0.29	
			P	6.29	4.57E-10	
			L	1.43	0.15	
	gpt-4-0125-preview	Sim. Sandín et al. [58] (199) Sim. Sandín et al. [58] 2001 (199)	E	0.35	0.73	396
			N	−1.20	0.23	
			P	−1.69	0.09	
			L	0.58	0.56	
		Sim. Sandín et al. [58] (199) Sim. Garcia et al. [24] (826)	E	−1.50	0.13	1,023
			N	3.06	2.24E-03	
			P	−0.16	0.88	
			L	−0.22	0.82	
		Sim. Sandín et al. [58] 2001 (199) Sim. Garcia et al. [24] (826)	E	−1.92	0.06	
			N	4.57	5.57E-06	
			P	1.99	0.05	
			L	−0.99	0.32	

Questionnaire in Spanish.

Table C7. Results of the *t*-Tests Comparing the Different Simulated Populations of Sandín et al. [58] and García-González et al. [24]

Language	Model	Population 1 (size) Population 2 (size)	Scale	T Statistics	p-value	Degrees of Freedom
English	gpt-3.5-turbo-0125	Sim. Sandín et al. [58] (199) Sim. Sandín et al. [58] 2001 (199)	E	−0.04	0.97	396
			N	−1.03	0.30	
			P	−2.24	0.03	
			L	−1.55	0.12	
		Sim. Sandín et al. [58] (199) Sim. Garcia et al. [24] (826)	E	−2.22	0.03	1,023
			N	2.08	0.04	
			P	−0.05	0.96	
			L	−1.01	0.31	
		Sim. Sandín et al. [58] 2001 (199) Sim. Garcia et al. [24] (826)	E	−2.10	0.04	
			N	3.48	5.27E-04	
			P	2.85	4.46E-03	
			L	0.95	0.34	
	gpt-4-0125-preview	Sim. Sandín et al. [58] (199) Sim. Sandín et al. [58] 2001 (199)	E	0.66	0.51	396
			N	−1.10	0.27	
			P	−1.18	0.24	
			L	0.14	0.89	
		Sim. Sandín et al. [58] (199) Sim. Garcia et al. [24] (826)	E	−2.64	8.42E-03	1,023
			N	3.41	6.69E-04	
			P	−2.15	0.03	
			L	0.09	0.93	
		Sim. Sandín et al. [58] 2001 (199) Sim. Garcia et al. [24] (826)	E	−3.18	1.50E-03	
			N	4.64	3.85E-06	
			P	−0.72	0.47	
			L	−0.08	0.93	

Questionnaire in English.

Table C8. Results of the *t*-Tests Comparing the Different Simulated Populations of Sandín et al. [58] and García-González et al. [24]

Language	Model	Population 1 (size) Population 2 (size)	Scale	T Statistics	p-value	Degrees of Freedom
Slovak	gpt-3.5-turbo-0125	Sim. Sandín et al. [58] (199) Sim. Sandín et al. [58] 2001 (199)	E	−1.90	0.06	396
			N	3.99	7.88E-05	
			P	−0.11	0.91	
			L	−2.50	0.01	
		Sim. Sandín et al. [58] (199) Sim. Garcia et al. [24] (826)	E	−1.67	0.09	1,023
			N	3.19	1.45E-03	
			P	−0.31	0.76	
			L	−1.53	0.13	
		Sim. Sandín et al. [58] 2001 (199) Sim. Garcia et al. [24] (826)	E	0.79	0.43	
			N	−1.75	0.08	
			P	−0.19	0.85	
			L	1.62	0.11	
	gpt-4-0125-preview	Sim. Sandín et al. [58] (199) Sim. Sandín et al. [58] 2001 (199)	E	0.06	0.95	396
			N	−2.24	0.03	
			P	0.95	0.34	
			L	0.83	0.40	
		Sim. Sandín et al. [58] (199) Sim. Garcia et al. [24] (826)	E	−2.12	0.03	1,023
			N	1.75	0.08	
			P	−0.74	0.46	
			L	1.93	0.05	
		Sim. Sandín et al. [58] 2001 (199) Sim. Garcia et al. [24] (826)	E	−2.24	0.03	
			N	4.57	5.49E-06	
			P	−1.96	0.05	
			L	0.95	0.34	

Questionnaire in Slovak.

Table C9. Results of the *t*-Tests Comparing the Different Results for Males and Females for Real and Simulated Populations

Gender	Model	Population 1 Population 2	Size	Personality Scale	T Statistics	p-value	Degrees of Freedom
Male	gpt-3.5-turbo-0125	Real Sandín [58] Sim. Sandín [58]	50	E	7.41	4.53E-11	98
				N	−8.19	1.01E-12	
				P	2.93	4.26E-03	
				L	−4.03	1.11E-04	
		Real Sandín [58] Sim. Sandín [58] 2001		E	8.13	1.37E-12	
				N	−10.74	2.97E-18	
				P	3.71	3.39E-04	
				L	−3.54	6.17E-04	
		Real Garcia [24] Sim. Garcia [24]	171	E	5.32	1.86E-07	340
				N	−6.77	5.61E-11	
				P	4.06	6.04E-05	
				L	−13.31	8.08E-33	
	gpt-4-0125-preview	Real Sandín [58] Sim. Sandín [58]	50	E	9.74	4.51E-16	98
				N	−8.10	1.53E-12	
				P	7.59	1.90E-11	
				L	2.30	0.023	
		Real Sandín [58] Sim. Sandín [58] 2001		E	9.66	6.55E-16	
				N	−6.08	2.27E-08	
				P	8.44	2.85E-13	
				L	2.58	0.011	
		Real Garcia [24] Sim Garcia [24]	171	E	6.11	2.75E-09	340
				N	−5.54	5.93E-08	
				P	13.50	1.50E-33	
				L	−5.10	5.55E-07	
Female	gpt-3.5-turbo-0125	Real Sandín [58] Sim. Sandín [58]	149	E	8.82	1.04E-16	296
				N	−12.79	4.15E-30	
				P	4.72	3.66E-06	
				L	−4.31	2.22E-05	
		Real Sandín [58] Sim. Sandín [58] 2001		E	8.95	3.90E-17	
				N	−12.65	1.34E-29	
				P	7.37	1.67E-12	
				L	−4.98	1.09E-06	
		Real Garcia [24] Sim. Garcia [24]	655	E	7.28	5.56E-13	1308
				N	−17.94	1.67E-64	
				P	8.35	1.71E-16	
				L	−10.73	8.26E-26	
	gpt-4-0125-preview	Real Sandín [58] Sim. Sandín [58]	149	E	10.53	3.14E-22	296
				N	−7.06	1.22E-11	
				P	12.18	5.95E-28	
				L	3.76	2.02E-04	
		Real Sandín [58] Sim. Sandín [58] 2001		E	9.91	3.54E-20	
				N	−6.41	5.86E-10	
				P	13.47	1.45E-32	
				L	3.07	2.33E-03	
		Real Garcia [24] Sim Garcia [24]	655	E	10.70	1.13E-25	1308
				N	−10.85	2.47E-26	
				P	24.98	0.00	
				L	12.50	5.73E-34	

Table C10. Results of the *t*-Tests Comparing the Different Results for Simulated Populations Aged 18–19 and 20+ for Spanish

Model	Population 1 (size) Population 2 (size)	Personality Scale	T Statistics	p-value	Degrees of Freedom
gpt-3.5-turbo-0125	Sandín [58] Age 18-19 (108) Age 20+ (91)	E	−0.12	0.91	197
		N	0.51	0.61	
		P	0.06	0.95	
		L	−0.36	0.72	
	Sandín [58] 2001 Age 18-19 (108) Age 20+ (91)	E	−0.17	0.87	
		N	0.72	0.47	
		P	0.77	0.44	
		L	1.94	0.05	
	Garcia [24] Age 18-19 (595) Age 20+ (231)	E	−1.57	0.12	824
		N	1.06	0.29	
		P	−0.44	0.66	
		L	3.68	2.50E-04	
gpt-4-0125-preview	Sandín [58] Age 18-19 (108) Age 20+ (91)	E	−0.21	0.84	197
		N	−0.07	0.94	
		P	1.13	0.26	
		L	−1.09	0.28	
	Sandín [58] 2001 Age 18-19 (108) Age 20+ (91)	E	−2.02	0.04	
		N	0.73	0.47	
		P	−0.36	0.72	
		L	−1.42	0.16	
	Garcia [24] Age 18-19 (595) Age 20+ (231)	E	−0.80	0.43	824
		N	1.34	0.18	
		P	−2.22	0.03	
		L	3.49	5.12E-04	

Table C11. Results of the *t*-Tests Comparing the Different Simulated Populations per Language

Model	Population 1 Population 2	Personality Scale	T Statistics	p-value	Degrees of freedom
gpt-3.5-turbo-0125	Sim. Sandín [58] Spanish Sim. Sandín [58] English	E	2.78	5.65E-03	396
		N	5.16	3.98E-07	
		P	0.55	0.58	
		L	7.06	7.36E-12	
	Sim. Sandín [58] Spanish Sim. Sandín [58] Slovak	E	−1.63	0.10	
		N	−9.07	5.77E-18	
		P	−2.44	0.02	
		L	−0.50	0.61	
	Sim. Sandín [58] 2001 Spanish Sim. Sandín [58] 2001 English	E	2.38	0.02	
		N	7.22	2.68E-12	
		P	−0.23	0.82	
		L	9.87	1.13E-20	
	Sim. Sandín [58] 2001 Spanish Sim. Sandín [58] 2001 Slovak	E	−3.04	2.49E-03	
		N	−6.58	1.50E-10	
		P	0.50	0.62	
		L	−3.68	2.70E-04	
	Sim. Garcia [24] Spanish Sim. Garcia [24] English	E	4.69	2.94E-06	1650
		N	10.19	1.14E-23	
		P	5.93	3.59E-09	
		L	18.87	4.77E-72	
	Sim. Garcia [24] Spanish Sim. Garcia [24] Slovak	E	−0.78	0.43	
		N	−16.73	3.54E-58	
		P	−10.52	4.33E-25	
		L	−6.03	2.00E-09	
gpt-4-0125-preview	Sim. Sandín [58] Spanish Sim. Sandín [58] English	E	1.38	0.17	396
		N	−2.22	0.03	
		P	−2.51	0.01	
		L	−8.61	1.81E-16	
	Sim. Sandín [58] Spanish Sim. Sandín [58] Slovak	E	0.54	0.59	
		N	−0.27	0.79	
		P	−3.37	8.13E-04	
		L	12.51	1.67E-30	
	Sim. Sandín [58] 2001 Spanish Sim. Sandín [58] 2001 English	E	1.07	0.29	
		N	−2.18	0.03	
		P	−3.49	5.30E-04	
		L	−8.42	7.18E-16	
	Sim. Sandín [58] 2001 Spanish Sim. Sandín [58] 2001 Slovak	E	0.24	0.81	
		N	−1.31	0.19	
		P	−0.66	0.51	
		L	13.72	2.48E-35	
	Sim. Garcia [24] Spanish Sim. Garcia [24] English	E	5.14	3.15E-07	1650
		N	−6.30	3.84E-10	
		P	−2.40	0.02	
		L	−18.99	7.51E-73	
	Sim. Garcia [24] Spanish Sim. Garcia [24] Slovak	E	0.41	0.68	
		N	−2.77	5.70E-03	
		P	−7.68	2.76E-14	
		L	32.94	0.00	

Table C12. Results of the t-Tests Comparing the Different Results for Males and Females for Real Populations

Paper	Population 1 (size) Population 2 (size)	Scale	T Statistics	p-value	Degrees of Freedom
Sandín et al. [58]	Female (149) Male (50)	E	1.80	0.07	197
		N	0.24	0.81	
		P	−0.77	0.44	
		L	−2.82	5.35E-03	
Garcia et al. [24]	Female (655) Male (171)	E	−0.47	0.64	824
		N	3.79	1.60E-04	
		P	−5.44	6.90E-08	
		L	−5.13	3.62E-07	

Table C13. Results of the *t*-Tests Comparing the Different Results for Males and Females for the Simulated Populations in Spanish

Model	Population 1 (size) Population 2 (size)	Scale	T Statistics	p-value	Degrees of Freedom
gpt-3.5-turbo-0125	Sim. Sandín et al. [58] Female (149) Male (50)	E	0.20	0.84	197
		N	0.75	0.45	
		P	−1.43	0.15	
		L	−1.67	0.10	
	Sim. Sandín et al. [58] 2001 Female (149) Male (50)	E	−0.74	0.46	
		N	3.18	1.71E-03	
		P	−0.61	0.54	
		L	−3.38	8.84E-04	
	Sim. Garcia et al. [24] Female (655) Male (171)	E	−2.90	3.84E-03	824
		N	0.72	0.47	
		P	−5.41	8.25E-08	
		L	4.98	7.82E-07	
gpt-4-0125-preview	Sim. Sandín et al. [58] Female (149) Male (50)	E	−0.76	0.45	197
		N	4.53	1.03E-05	
		P	−2.80	5.56E-03	
		L	−3.49	6.01E-04	
	Sim. Sandín et al. [58] 2001 Female (149) Male (50)	E	−1.38	0.17	
		N	2.70	7.63E-03	
		P	−2.38	0.02	
		L	−4.78	3.39E-06	
	Sim. Garcia et al. [24] Female (655) Male (171)	E	−0.84	0.40	824
		N	3.98	7.45E-05	
		P	−4.08	4.92E-05	
		L	8.98	1.83E-18	

Received 11 March 2024; revised 1 August 2024; accepted 11 December 2024