Generating Social and Emotional Skill Items: Humans vs. ChatGPT

Kate E. Walton and Cristina Anguiano-Carrasco

Large language models (LLMs), such as ChatGPT, are becoming increasingly prominent. Their use is becoming more and more popular to assist with simple tasks, such as summarizing documents, translating languages, rephrasing sentences, or answering questions. Reports like McKinsey's (Chui, & Yee, 2023) estimate that by implementing LLMs, corporations could see a potential growth of $4.4 trillion annually in corporate benefits, while Nielsen (2023) estimates a 66% increase in employee productivity when using LLMs and other forms of generative artificial intelligence (AI). Can we use ChatGPT in the field of social and emotional learning assessment development to enhance our productivity?

Some have examined how social and emotional (SE) skills are related to ChatGPT usage, such as cheating in the academic domain (Greitemeyer & Kastenmüller, 2023). In another study, researchers (de Winter et al., 2023) had ChatGPT generate a large number of personas and complete several SE skill measures. They then carried out several analyses such as a factor analysis and correlations with outcome measures and determined how similar the results were to previous research using human-completed SE skill measures. In the current study, rather than have ChatGPT complete SE skill measures, we sought to have ChatGPT create SE skill measures. Ultimately, we will compare a ChatGPT-generated assessment with a human-generated assessment in terms of reliability and validity.

The Current Study

Phase 1: Item Generation

We gave ChatGPT and two human item writers the same instructions to write Likert and forced choice (FC) items to cover the domains of conscientiousness (i.e., the tendency to be persistent, reliable, dependable, etc.), agreeableness (i.e., the tendency to be empathic, helpful, trustworthy, etc.), and emotional stability (i.e., the tendency to be stress tolerant, calm, poised, etc.). The humans were PhD-level subject matter experts (SMEs). An example of a Likert item is a statement such as, *I check my work before turning it in* (an indicator of conscientiousness). Respondents indicate on a scale of 1–6 how much they agree or disagree with the statement. An example of an FC triad is three statements such as, *I check my work before turning it in*, *I am a good team player* (an indicator of agreeableness), and *I can handle stress well* (an indicator of emotional stability). Respondents select the item that is most like them and least like them.

Simply, the instructions for generating Likert items were: *Create six Likert conscientiousness items.* The human writers and ChatGPT were then asked to create six agreeableness items and six emotional stability items. The instructions for generating FC items were: *Create five* *multidimensional forced choice triads to measure conscientiousness, agreeableness, and emotional stability.* Sample items created can be found in Table 1.

Table 1. Sample Items Written by Humans and ChatGPT

	Item	Intended Skill
Human-Generated Likert Items	I always have my school materials organized so I can easily find them.	Conscientiousness
	It's hard for me to get along with some people.	Agreeableness
	It's easy for me to move on from sadness.	Emotional Stability
ChatGPT-Generated Likert Items	I find it challenging to relax until I've completed all of my tasks for the day.	Conscientiousness
	I believe that cooperation and teamwork lead to better outcomes than individual efforts.	Agreeableness
	I can stay composed and rational even in highly stressful situations.	Emotional Stability
Human-Generated Forced Choice Items	I follow through on my commitments.	Conscientiousness
	I like to give others compliments.	Agreeableness
	I often have intense emotions.	Emotional Stability
ChatGPT-Generated Forced Choice Items	I am known for being responsible and dependable in both my personal and professional relationships.	Conscientiousness
	I value harmony and cooperation, always striving to maintain positive interactions with others.	Agreeableness
	I handle stressful situations with composure and focus, ensuring that tasks are completed efficiently.	Emotional Stability

Phase 2: Preliminary Observations and Analyses

A few things were immediately apparent upon reflection. First, the human writers generated some reverse-keyed items (e.g., *I often have intense emotions*), while ChatGPT did not. The item provided as an example is meant to be an indicator of emotional stability, but it is written in the direction of low emotional stability. Second, ChatGPT included double-barreled items, items that include more than one sentiment. For example, the item *I am known for being responsible and dependable in both my personal and professional relationships* is double-barreled; it asks not only about being both responsible and dependable, but also about both one's personal and professional relationships. Third, ChatGPT generated some multidimensional items. For example, *I handle stressful situations with composure and focus, ensuring that tasks are completed efficiently* is multidimensional, tapping into emotional stability (*I handle stressful situations with composure and focus*) as well as conscientiousness (*ensuring that tasks are completed efficiently*). Finally, it is clear that ChatGPT's items are longer and have a greater cognitive load. The average length of the human-generated Likert items is 7.6 words with a Flesch-Kincaid reading level of 4.1. The average length of the ChatGPT-generated Likert items is 12.6 words with a Flesch-Kincaid reading level of 10.1. The average length of the human-generated FC items is 6.4 words with a Flesch-Kincaid reading level of 4.7. The average length of the ChatGPT-generated Likert items is 14.8 words with a Flesch-Kincaid reading level of 13.2.

Phase 3: Subject Matter Expert Ratings

We solicited input from four SMEs who were PhD-level researchers in social and emotional learning. We first asked them to rate each human- and ChatGPT-generated item on the following: *How good of an indicator is this item for its intended skill?* They rated each item on a scale of 1 (*very bad*) to 6 (*very good*). They were blind to which items came from which source. There was no significant difference between the human-generated (M = 5.28, SD = .67) and ChatGPT-generated (M = 5.40, SD = .54) Likert items, t = −.62, p = .54. The effect size was d = −.21. The human-generated FC items (M = 5.47, SD = .67) were rated as better indicators than the ChatGPT-generated FC items (M = 3.97, SD = 1.35), t = 3.86, p < .01. The effect size was large, d = 1.41.

We then asked the SMEs to rate each human- and ChatGPT-generated item on the following: *How natural does this item's language sound?* They rated each item on a scale of 1 (*very unnatural*) to 6 (*very natural*). There was no significant difference between the human-generated (M = 4.78, SD = .93) and ChatGPT-generated (M = 5.06, SD = 1.06) Likert items, t = −.84, p = .41. The effect size was d = −.28. There was no significant difference between the human-generated (M = 5.12, SD = .87) and ChatGPT-generated (M = 4.93, SD = 1.08) FC items, t = .51, p = .61. The effect size was d = .19.

Phase 4: Student Survey

Method

We sought participation from students who took the ACT® test on the September 2023 national test date. An invitation went out to a random sample of 30,000 students inviting them to participate in research. They were not incentivized to participate, and they were assured that their involvement and responses would not impact their ACT scores.

We have complete data for 1,707 participants. Of the sample, 1,198 (70.2%) identified as female, 474 (27.8%) identified as male, seven (.4%) identified as another gender, 27 (1.6%) preferred to not respond, and the information was missing for one participant. Additionally, 1,130 (66.2%) of the sample identified as White, 182 (10.7%) identified as Asian, 137 (8.0%) identified as Hispanic/Latino, 94 (5.5%) identified as Black/African American, 79 (4.6%) identified as two or more races, one (.1%) identified as American Indian/Alaska Native, 79 (4.6%) preferred to not respond, and the information was missing for one participant. Two (.1%) of the students were in 8th grade, nine (.5%) were in 9th grade, 82 (4.8%) were in 10th grade, 662 (38.8%) were in 11th grade, 922 (54.0%) were in 12th grade, eight (.5%) were college students, and the information was missing for 22 participants.

Participants completed either the human- (n = 919) or ChatGPT-generated (n = 788) assessments. All participants completed the test-criterion validity measure (items can be found in Table 3). There were three items that should correlate most highly with conscientiousness, there were two items that should correlate most highly with agreeableness, and there were two items that should correlate most highly with emotional stability.

Results

Reliability

We first calculated Cronbach's alpha values for each human- and ChatGPT-generated scale (see Table 2). For the Likert items, the values across humans and ChatGPT were relatively similar, and in all but one case (human-generated conscientiousness items), alpha exceeded .70, which is the standard mark of acceptable reliability. For the FC items, alpha values were generally lower, which is typical given the ipsative nature of the scales. After removal of some problematic items, the ChatGPT scales had higher reliability estimates than the human-generated scales.

Table 2. Cronbach's Alpha Values

Skill	Likert		Forced Choice	
	Human	ChatGPT	Human	ChatGPT
Conscientiousness	.68	.75	.45	.47[b]
Agreeableness	.76	.74	.54	.67[c]
Emotional Stability	.80	.81	.40[a]	.44[d]

Note. The reliability of some scales could be improved by the removal of one item. The following are the alpha values after removal of one item: [a].63, [b].63, [c].79, [d].72.

Validity

We next evaluated the structural validity of the two Likert assessments. A three-factor confirmatory factor model was fit to the data. The human-generated assessment had reasonable fit, CFI = .86, TLI = .84, RMSEA = .11. The ChatGPT-generated assessment had slightly better fit, CFI = .90, TLI = .88, RMSEA = .08. However, the inter-factor correlations were higher for the ChatGPT-generated assessment; the average correlation was .47 for ChatGPT and .39 for humans.

Finally, we examined correlations between the skills and the test-criterion validity measure. Here we call out any differences between correlations that reach .10. Both the human- and ChatGPT-generated assessments show evidence of test-criterion validity. For Likert items (Table 3), ChatGPT had stronger evidence with the correlations between 1) conscientiousness and challenging the self to work harder, 2) agreeableness and getting along with others who are different, and 3) agreeableness and being respectful of others who disagree. The human-generated assessment, however, had stronger evidence for emotional stability and its correlation with the number of days in the past week feeling nervous. For the FC assessment (Table 4), the human-generated conscientiousness scale had a stronger correlation with challenging oneself to work harder.

Table 3. Likert Scales’ Correlations with Test-Criterion Validity Variables

Outcome	Conscientiousness		Agreeableness		Emotional Stability	
	Human	ChatGPT	Human	ChatGPT	Human	ChatGPT
Grade point average	.25*	.26*	−.03	−.03	−.04	−.03
Challenging oneself to work harder	.33*	.51*	.11	.30	.11	.25
Checking that homework is free from errors before turning in	.35*	.43*	.13	.21	.06	.15
Getting along with others who are different	.09	.16	.28*	.46*	.13	.21
Being respectful of others who disagree	.17	.17	.25*	.38*	.22	.29
Allowing setbacks to affect mood for the day	−.19	.06	−.20	−.05	−.51*	−.46*
Days in past week feeling nervous	−.07	.06	−.14	.04	−.44*	−.34*

Note. *Indicates correlations that should be highest, suggesting evidence of test-criterion validity.

Table 4. Forced Choice Scales' Correlations with Test-Criterion Validity Variables

Outcome	Conscientiousness		Agreeableness		Emotional Stability	
	Human	ChatGPT	Human	ChatGPT	Human	ChatGPT
Grade point average	.21*	.20*	.12	.14	.05	.03
Challenging oneself to work harder	.30*	.20*	−.13	−.15	−.12	−.01
Checking that homework is free from errors before turning in	.28*	.22*	−.05	−.10	−.20	−.09
Getting along with others who are different	−.03	−.09	.16*	.20*	−.09	−.13
Being respectful of others who disagree	.06	−.08	.04*	.12*	−.04	−.06
Allowing setbacks to affect mood for the day	.02	.14	−.12	.14	−.30*	−.27*
Days in past week feeling nervous	.02	.06	−.13	.14	−.27*	−.21*

Note. *Indicates correlations that should be highest, suggesting evidence of test-criterion validity.

Conclusion

To our knowledge, this is the first attempt made to compare a ChatGPT-generated SE skills assessment with a traditional human-generated one. ChatGPT violated some basic item writing guidelines like including double-barreled items and writing unnecessarily long item stems with a high reading load. In addition, ChatGPT generated some items that were multidimensional, which presumably is why the SME rated them as poorer indicators of their intended skill. Moreover, this likely explains why the confirmatory factor model for the ChatGPT assessment had stronger inter-factor correlations than the human-made assessment. ChatGPT, however, had stronger evidence of internal consistency reliability, particularly for the FC assessment. As far as test-criterion validity, both assessments had sound validity evidence, and each outperformed the other in certain instances.

In sum, it seems that ChatGPT is a viable resource for generating SE skill items. With additional prompts (e.g., about reading level, avoiding double-barreled items, etc.), ChatGPT would likely perform even better than what we observed here. However, we would caution anyone against blindly using ChatGPT to generate assessments. ChatGPT can be wrong even when asked simple math questions. One study (Chen et al., 2023) showed that over the course of a few months, ChatGPT went from correctly answering a simple math problem 98% of the time to a mere 2% of the time. The interface itself cautions against using it blindly, reading, “ChatGPT can make mistakes. Consider checking important information” (ChatGPT, 2024). Instead of relying solely on ChatGPT, we would argue that it should be used as a supplementary tool for assessment item generation.

References

ChatGPT. (2024, February 15). *ChatGPT 3.5*. https://chat.openai.com/

Chen, L., Zaharia, M., & Zou, J. (2023). *How is ChatGPT’s behavior changing over time?* [Unpublished manuscript]. https://arxiv.org/pdf/2307.09009.pdf

Chui, M., & Yee, L (2023, July 7). *AI could increase corporate profits by $4.4 trillion a year, according to new research*. https://www.mckinsey.com/mgi/overview/in-the-news/ai-could-increase-corporate-profits-by-4-trillion-a-year-according-to-new-research#

de Winter, J. C. F., Driessen, T., & Dodou, D. (2023). *The use of ChatGPT for personality research: Administering questionnaires using generated personas.* [Unpublished manuscript]. https://www.researchgate.net/publication/374415968_The_use_of_ChatGPT_for_personality_research_Administering_questionnaires_using_generated_personas

Greitemeyer, T., & Kastenmüller, A. (2023). HEXACO, the Dark Triad, and Chat GPT: Who is willing to commit academic cheating? *Heliyon*, *9*(9).

Nielsen, J. (2023, July 16). *AI improves employee productivity by 66%.* https://www.nngroup.com/articles/ai-tools-productivity-gains/

ABOUT ACT

ACT is a mission-driven, nonprofit organization dedicated to helping people achieve education and workplace success. Grounded in more than 60 years of research, ACT is a trusted leader in college and career readiness solutions. Each year, ACT serves millions of students, job seekers, schools, government agencies, and employers in the U.S. and around the world with learning resources, assessments, research, and credentials designed to help them succeed from elementary school through career.

For more information, visit act.org