中国心理学前沿
2024 年 3 月第 6 卷第 3 期

基于 ChatGPT 的中文版情绪稳定性问卷的开发探索与信效度验证

高垚杰

天津师范大学心理学部，天津

摘　要 | 随着情绪稳定性越来越受到社会与研究者的广泛关注，并鉴于其对于国民健康及多个研究领域的重要作用，质量良好的情绪稳定性问卷显得尤为关键。但目前情绪稳定性的测量多借助于大五人格量表中的神经质维度，既不利于大众的理解与传播，也揭露了本土化测量工具有待丰富的现状，因此本研究旨在开发中文版的情绪稳定性问卷，并顺应人工智能的发展潮流，学习国外最新的项目开发范式，探索利用ChatGPT生成中文版情绪稳定性项目的可行性，并通过专家筛选、项目分析、信效度验证等步骤，验证GPT生成项目的性能，以期形成一个质量良好的中文版情绪稳定性问卷。

关键词 | ChatGPT；问卷开发；情绪稳定性；信度；效度

1　引言

随着“emo”“网抑云”等蕴含沮丧情绪的网络热词的流行，如何保持情绪稳定越来越受到社会的广泛关注。2020 年中国青年报社社会调查中心联合问卷网对 3006 名受访者的一项调查显示，76.6% 的受访者觉得做个情绪稳定的人很困难。在这样的大环境下，使用专业的心理学测量工具对情绪稳定性进行调查，从而筛选出情绪不稳定人群，以针对性进行预防甚至干预，对于促进国民心理健康是十分必要的。

同时情绪稳定性也受到了研究者的青睐，通过在 web of science 以情绪稳定性为主题检索发现，自 2019 年以来共有相关文章 2163 篇，篇均被引频次 7.89 次。在过去的几年间，由于疾病的威胁、防控的要求、政治局势的紧张、经济的动荡等因素，包括抑郁、失眠、恐惧和焦虑在内的心理健康问题有所增加[1]。Margetić 等人提出情绪稳定性是心理健康最重要的预测指标[2]，情绪不稳定的人更容易经历负面情绪从而导致更多不良后果。而情绪不稳定对个体带来的负面影响涵盖了多个领域，如在组织管理领域的研究

基金项目：国家级大学生创新创业训练计划项目“情绪稳定性的计算机化自适应性测验系统的开发”（项目编号：202310065022）。
作者简介：高垚杰，天津师范大学应用心理学（创新班）本科在读，研究方向：心理测量。
文章引用：高垚杰. 基于ChatGPT的中文版情绪稳定性问卷的开发探索与信效度验证 [J]. 中国心理学前沿，2024，6（3）：328-336.
https://doi.org/10.35534/pc.0603037

中，发现情绪稳定性低会带来更高程度的情绪耗竭，并对创造力有负面影响[3, 4]；在心理健康领域中，情绪稳定性低会给个人幸福感带来负面影响[5]；在教育背景下的研究中发现，班主任的低情绪稳定性会影响其对师生关系和课堂管理技能做出更消极的评价[6]；在生理心理学中，有研究者发现情绪不稳定会使睡眠质量变差[7]；在决策的研究中，有学者提出较低的情绪稳定性与放大风险感知、更多的冒险行为有关[8, 9]。而以上种种研究，同样都离不开信效度高的情绪稳定性测量工具。

综上，不难发现情绪稳定性在现实与理论研究中都有重要意义。至于情绪稳定性是什么，目前学者们仍持有不同的见解。国外学者雷柏（Raber）在 *The Penguin Dictionary of Psychology* 中定义情绪稳定性为“用来描述一个人在情感上的成熟状态，情绪稳定的人面对不同的情境会做出合适且较为一致的情绪反应”[10]。朱智贤在《心理学大辞典》则定义为：“人的情绪状态受外界（或内部）条件而产生波动的情况。一些情绪较为稳定的人不易为一般的情境引起强烈的情绪反应，或引起的情绪反应较为缓慢，情绪不稳定的人对事物的发生则容易引起情绪反应，生活琐碎小事也可招致强烈的情绪变化，一经引起情绪波动，对情绪的控制较差。这种情绪的稳定性与个人意志强弱有关，是人格的重要特质之一。”[11] 张春兴在《张氏心理学大辞典》中认为“情绪稳定性是一种人格特质，使个体在一定的情绪情景中不表现出过激的反应”[12]。

虽然定义尚未统一，但把情绪稳定性作为人格特质已是当前研究者的共识。但如今情绪稳定性的测量常常借助于大五人格模型中的神经质（Neuroticism）维度，两者是处在一个量尺上的两端的关系[13]，但神经质这一概念不如情绪稳定性便于大众理解与传播。此外，回顾已有测量工具不难发现，本土化的大五人格量表有待进一步丰富。因此，本研究旨在开发中文版的情绪稳定性测量工具。

随着人工智能（artificial intelligence，AI）以及自然语言处理（natural language processing，NLP）的高速发展，已有国外学者利用生成式预训练模型（generative pre-trained，GPT）进行大五人格项目的自动项目生成（automatic item generation，AIG）并验证了其性能，研究结果很大程度上肯定了该技术在项目开发上的经济性与便利性[14, 15, 16]。随着后续 OpenAI 发布的 ChatGPT，因其对话式的使用模式，不再需要扎实的编程与自然语言处理基础，极大降低了上手门槛。同时基于 GPT-4 的 ChatGPT 展现出了优异的中文性能与丰富的心理学知识[17]，尽管如此，目前国内缺乏相关的项目开发的探索。因此本研究意欲借助于最新版本的 ChatGPT，进行情绪稳定性的项目开发与问卷信效度验证的探索，以期形成一个质量良好的中文版情绪稳定性测量工具。

2 研究方法

2.1 项目初步开发流程

2.1.1 项目生成

为了让 ChatGPT 生成符合要求的项目，我们进行了一系列中文的提示工程（Prompt Engineering）：（1）让 GPT 扮演项目开发经验丰富的专家；（2）为其介绍上述国内外知名学者对于情绪稳定性的定义；（3）介绍情绪稳定性作为一种人格特质常常通过大五人格测验中的神经质子维度进行测量，由此向 GPT 提出参考已有项目来制定全新的情绪稳定性项目这一任务；（4）为 GPT 详细规定制定项目过程中

https://doi.org/10.35534/pc.0603037

必须遵循的基本原则，包括原创性、明确测量目标、丰富项目形式、多样性与全面性等。尤其最后一点值得注意，该原则源自前人研究提出的 GPT 生成项目存在情境上过于相似的问题[16]，因此在阐述该规则时，研究者应当为 GPT 提供与测验目标相关的尽可能多的情境及相关因素作为辅助，这些解释有助于提高生成项目的质量[18]；（5）最后，要求 GPT 根据提供的每组已有的神经质项目作为参考制定项目。

2.1.2 项目初步筛选

根据上述步骤，我们保留了 GPT 制定的计分方式相同的 104 道测量情绪稳定性的项目，所有项目采用李克特 5 点计分。然后邀请 10 位心理学专业的研究生对项目的语法与内容有效性进行判断。根据结果，首先 1 道表述不当的项目被剔除。接着 10 位专家对每个项目代表情绪稳定性这一人格特质的程度进行 4 点评分。根据评分结果，计算了修正后的 Kappa 系数（k*），结果如图 1 所示，按照评估标准，k* 大于 0.74 的属于优质项目[19]，因此所有不符合该标准的项目被剔除，最终 68 道项目得到保留以进行正式施测。

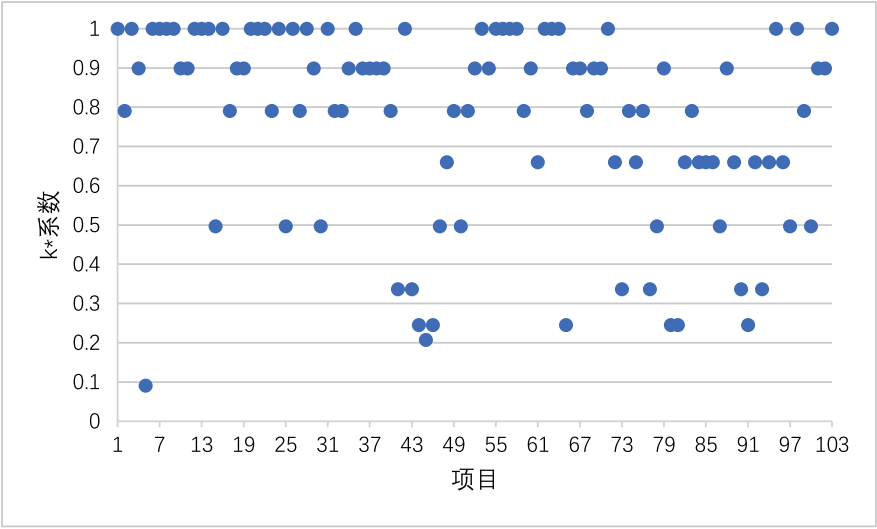

图 1 k* 指数

Figure 1 k* index

2.2 项目检验

2.2.1 被试

本研究采用完全随机抽样，采用问卷星的网络测验方式，共获得有效样本 479 人（其中男性 163 人），平均年龄为 22.82 岁（SD=6.52），所有被试均已知情同意。

2.2.2 研究工具

除上述 68 道 GPT 生成的项目，我们另外选择了 4 个已被广泛使用的大五人格量表中共计 42 道测量情绪稳定性（神经质）的项目加入施测项目中，用作后续的性能对比与校标量表。其中包含中国大五人格问卷简式版的 8 题（CBF–PI–B）[20]，BFI–2 中文版的 12 题[21]，翻译后的 TIPI–10 的 2 题[22] 与 IPIP–BFAS 的 20 题[23]。在本研究中，这四个量表的情绪稳定性维度的 Cronbach's α 分别为 0.921，0.935，0.799，0.955。

本研究中所有需要反向计分的项目均已进行相关处理，所有项目得分越高代表情绪稳定性水平越高。

2.2.3 数据分析

使用 SPSS 26.0 实现对数据的录入、描述统计、项目分析、探索性因素分析、校标关联效度与信度的计算；利用 Mplus 8.0 进行验证性因素分析。

3 研究结果

3.1 项目分析

对 68 道 GPT 生成的情绪稳定性项目进行区分度计算，结果发现所有项目区分度均大于 0.4（M=0.75，SD=0.08）。为进一步确认题目鉴别不同被试情绪稳定性的能力，我们计算了临界比率（critical ratio，CR），具体来说，是将被试在所有项目上作答的总分从高到低排序，将总分最高的 27% 的划为高分组，最低的 27% 的划为低分组，通过对两组被试在各个题目上的得分进行独立样本 t 检验来判断该项目的区分度水平，如果结果不显著代表项目鉴别力不足。经检验，GPT 生成的 68 道项目均有良好的区分度（p<0.001）。

接着计算题总相关系数（Item–total Correlation），结果发现所有项目与总分的相关系数均大于 0.4（p<0.001）。因而 68 道项目均得以保留，对其进行难度分析，结果如图 2 所示，可见项目难度处于 0.61~0.78 的范围区间，整体上难度偏低。

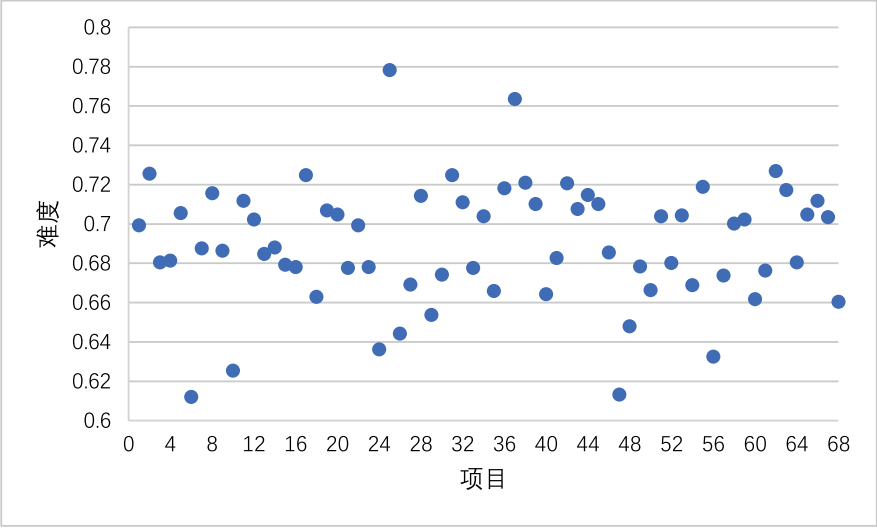

图 2 GPT 生成项目的难度值

Figure 2 Difficulty of GPT-generated items

3.2 效度分析

3.2.1 探索性因素分析

对 68 道题初步分析得到 KMO 值为 0.986（χ^2（2278）= 31922.37，p < 0.001），表明其适合进行因素分析。接着采用主成分分析和最优斜交法对量表的结构进行探索性因素分析。从碎石图（图 3）可以发现在第一因子后便出现明显拐点，结合第一因子累积方差解释率高达 58.11% 与事先理论构想，最终

https://doi.org/10.35534/pc.0603037

确立所有项目符合单因子结构。

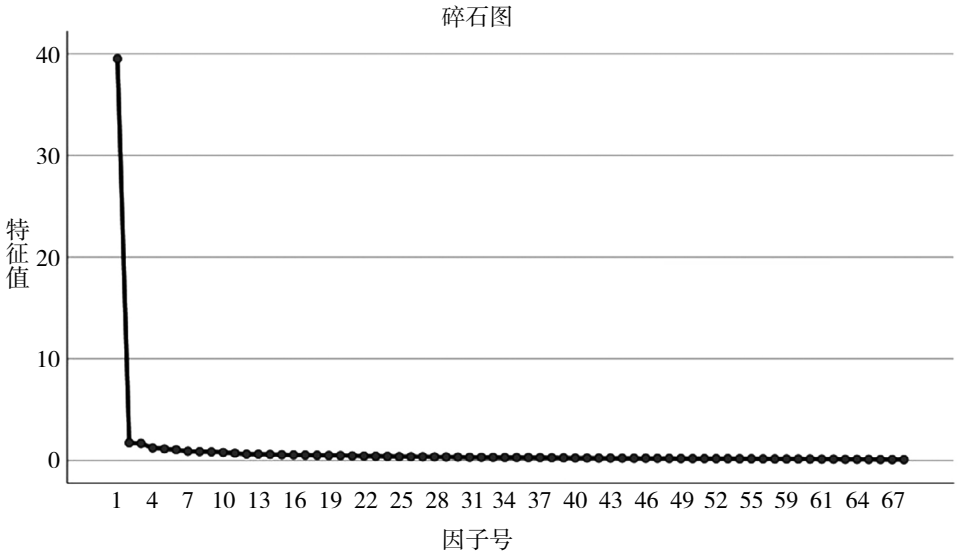

图 3　碎石图

Figure 3　Gravel diagram

为避免项目数过多带来的测试效率降低，对现有项目进行筛选，采用主轴因式分解提取 1 个因子，在剔除语义过于相似的项目后，选择其中共同度最高的 15 题。此时对保留的项目再次进行分析，此时 KMO 值为 0.978（χ^2（105）= 6997.35，$p < 0.001$），同样采用主成分分析和最优斜交法，发现第一因子累积方差解释率高达 71.05%，依旧符合单因子结构，接着采用主轴因式分解提取 1 个因子，得到所有项目共同度均高于 0.6，且因子载荷均达到 0.78 以上，详细数据如表 1 所示，表明指标良好。

表 1　15 个项目因子载荷与共同度

Table 1　Factor loadings and commonality of the 15 items

项目	第一因子载荷	共同度
1. 在繁忙或压力较大的日子里，我能保持心态平和	0.817	0.667
2. 我能够承受生活中的重大事件和压力	0.843	0.710
3. 我能在压力环境中保持冷静和专注	0.882	0.778
4. 我能有效应对日常的压力和挑战	0.842	0.710
5. 我通常保持乐观，很少感到忧郁	0.813	0.661
6. 在遭遇意外困难时，我能够迅速调整情绪	0.861	0.741
7. 我能在日常生活中有效处理压力	0.842	0.709
8. 在紧张或困难的情况下，我能保持乐观态度	0.830	0.689
9. 在遭遇失败时，我通常能保持积极和乐观	0.815	0.664
10. 面对紧急情况时，我通常能保持镇定	0.843	0.711
11. 我通常能在情绪波动时保持冷静和理智	0.837	0.700
12. 我在各种情况下都能保持情绪的稳定	0.828	0.686
13. 我很少因突发变化而感到不安或焦虑	0.803	0.646
14. 我通常能有效地控制和管理自己的情绪	0.789	0.623
15. 在听到不利消息时，我能够保持冷静	0.809	0.655

3.2.2 验证性因素分析

对保留的 15 个项目进行单因子的验证性因素分析。一般认为 TLI、CFI>0.90，RMSEA<0.08，SRMR<0.08 表示模型拟合良好[24]，从表 2 的结果可以发现 15 个项目拟合指标良好，问卷具有良好的结构效度。

表 2 验证性因素分析结果

Table 2 Results of confirmatory factor analysis

χ^2	df	χ^2/df	TLI	CFI	RMSEA	SRMR
330.408	90	3.67	0.960	0.966	0.075	0.023

3.2.3 校标关联效度分析

为进一步验证 GPT 生成项目的效度，我们以四个经典量表的神经质维度作为校标量表计算校标关联效度，结果（见表 3）表明由 GPT 生成项目组成的问卷效标效度良好。

表 3 校标关联效度

Table 3 Criterion-related validity

	CBF-PI-B	中文版 BFI-2	TIPI-10	IPIP-BFAS
GPT 问卷	0.755***	0.846***	0.811***	0.847***

注：*** 表示 $p<0.001$。

3.3 信度分析

对保留的 15 道 GPT 项目组成的问卷进行信度分析，并与四个经典量表的结果进行对比，具体内容包括 Cronbach's α 系数和分半信度，结果如表 4 所示。从中不难发现 GPT 问卷信度良好，且优于经典量表中的神经质子维度。

表 4 信度分析结果

Table 4 Results of the reliability analysis

问卷	Cronbach's α	分半信度	题目数
GPT 问卷	0.971	0.965	15
CBF-PI-B	0.921	0.906	8
中文版 BFI-2	0.935	0.932	12
TIPI-10	0.799	0.801	2
IPIP-BFAS	0.955	0.874	20

4 讨论

鉴于情绪稳定性的重要性及相关本土化测量工具的缺乏，本研究紧跟 AI 的时代潮流，学习与参考

https://doi.org/10.35534/pc.0603037

国外最新项目开发范式，利用基于 GPT-4 Turbo 的 ChatGPT 按标准化流程制定了大量情绪稳定性项目，并按照一系列筛选与检验流程，验证了人工智能生成中文版情绪稳定性项目的可行性，最终构建了一个信效度良好的情绪稳定性问卷。

传统的问卷开发需要专家利用其丰富的经验与知识来构建新项目，并不断审查、修改与完善，直至满足所需的质量标准，而这一过程会耗费大量人力、物力与财力[25]，而本研究的方法为项目开发提供了一种高效且经济的途径，相较于前人所使用的旧版本 GPT[15, 16, 26]，最新的 ChatGPT 不再要求研究者拥有扎实的机器学习基础以针对项目开发任务微调模型，极大降低了项目开发者的上手门槛，并且展现了更优异的性能。

尽管本研究为最新的自然语言处理技术与情绪稳定性项目开发的结合提供了有力证据，但仍然存在一些局限之处：首先为了避免样本造成的偏差，后续还有待对问卷进行跨样本的验证；其次，GPT 生成的项目仍存在一些不足，如缺乏反向计分的项目，整体难度较低，问题所涉及的相关情境有限等。

在后续的研究中，GPT 能否应用于更广泛测量目标的项目生成需要进一步探索，尤其是针对一些缺乏可靠测量工具的目标概念。另外除了经典的李克特式计分项目，迫选题等更丰富的项目形式能否借助 GPT 得到进一步发展也令人十分期待。在具体的实际应用中，基于 ChatGPT 的自动项目生成被认为有望推动计算机化自适应性测验的发展[14]，当前自适应测验这一智能化的测验方式受限于其所需的大型题库开发的高昂成本而难以普及[27, 28]，而使用 ChatGPT 生成项目的经济性与高效性很大程度上解决了这一痛点。

所有受试者在参与研究前均已知情同意纳入研究。该研究是根据《赫尔辛基宣言》进行的，该议定书获得了天津师范大学伦理委员会（2023080902）的批准。

参考文献

[1] Courtney D，Watson P，Battaglia M，et al. COVID-19 Impacts on Child and Youth Anxiety and Depression：Challenges and Opportunities [J]. The Canadian Journal of Psychiatry，2020，65（10）：688-691.

[2] Margetić B，Peraica T，Stojanović K，et al. Spirituality，Personality，and Emotional Distress During COVID-19 Pandemic in Croatia [J]. Journal of Religion and Health，2022，61（1）：644-656.

[3] David E M，Shoss M K，Johnson L U，et al. Emotions running high：Examining the effects of supervisor and subordinate emotional stability on emotional exhaustion [J]. Journal of Research in Personality，2020（84）：103885.

[4] Park I-J，Shim S-H，Hai S，et al. Cool down emotion，don't be fickle! The role of paradoxical leadership in the relationship between emotional stability and creativity [J]. The International Journal of Human Resource Management，2022，33（14）：2856-2886.

[5] Kobyliń ska D，Zajenkowski M，Lewczuk K，et al. The mediational role of emotion regulation in the relationship between personality and subjective well-being [J]. Current Psychology，2022，41（6）：4098-4111.

[6] Wettstein A，Ramseier E，Scherzinger M. Class- and subject teachers' self-efficacy and emotional stability and students' perceptions of the teacher-student relationship，classroom management，and classroom disruptions [J]. BMC Psychology，2021，9（1）：103.

[7] Lau E Y Y, Li C, Hui C H, et al. A longitudinal investigation of the bidirectional relationship of sleep quality with emotional stability and social cynicism in a large community sample [J]. Sleep Health, 2021, 7(3): 384–389.
[8] Bec A, Becken S. Risk perceptions and emotional stability in response to Cyclone Debbie: An analysis of Twitter data [J]. Journal of Risk Research, 2021, 24(6): 721–739.
[9] Zhang Q, Wang X, Miao L, et al. The Effect of Chronotype on Risk–Taking Behavior: The Chain Mediation Role of Self–Control and Emotional Stability [J]. International Journal of Environmental Research and Public Health, 2022, 19(23): 16068.
[10] Reber A S. The Penguin dictionary of psychology [M]. Penguin Press, 1995.
[11] 朱智贤. 心理学大词典 [M]. 北京师范大学出版社, 1989.
[12] 张春兴. 张氏心理学辞典 [M]. 上海辞书出版社, 1992.
[13] Costa P T, McCrae R R. Normal personality assessment in clinical practice: The NEO Personality Inventory [J]. Psychological assessment, 1992, 4(1): 5.
[14] Hommel B E, Wollang F–J M, Kotova V, et al. Transformer–Based Deep Neural Language Modeling for Construct–Specific Automatic Item Generation [J]. Psychometrika, 2022, 87(2): 749–772.
[15] Götz F M, Maertens R, Loomba S, et al. Let the algorithm speak: How to use neural networks for automatic item generation in psychological scale development [M]. Psychological Methods, 2023.
[16] Lee P, Fyffe S, Son M, et al. A Paradigm Shift from "Human Writing" to "Machine Generation" in Personality Test Development: An Application of State–of–the–Art Natural Language Processing [J]. Journal of Business and Psychology, 2023, 38(1): 163–190.
[17] Achiam J, Adler S, Agarwal S, et al. Gpt–4 technical report [R]. arXiv preprint arXiv: 2303. 08774.
[18] Lampinen A K, Dasgupta I, Chan S C Y, et al. Can language models learn from explanations in context? (arXiv: 2204. 02329). arXiv [EB/OL]. [2024–03–18]. http://arxiv.org/abs/2204.02329.
[19] Polit D F, Beck C T, Owen S V. Is the CVI an acceptable indicator of content validity? Appraisal and recommendations [J]. Research in Nursing & Health, 2007, 30(4): 459–467.
[20] 王孟成, 戴晓阳, 姚树桥. 中国大五人格问卷的初步编制Ⅲ: 简式版的制定及信效度检验 [J]. 中国临床心理学杂志, 2011(4): 454–457.
[21] Zhang B, Li Y M, Li J, et al. The Big Five Inventory–2 in China: A Comprehensive Psychometric Evaluation in Four Diverse Samples [J]. Assessment, 2022, 29(6): 1262–1284.
[22] Gosling S D, Rentfrow P J, Swann W B. A very brief measure of the Big–Five personality domains [J]. Journal of Research in Personality, 2003, 37(6): 504–528.
[23] DeYoung C G, Quilty L C, Peterson J B. Between facets and domains: 10 aspects of the Big Five [J]. Journal of Personality and Social Psychology, 2007, 93(5): 880–896.
[24] 温忠麟, 侯杰泰, 马什赫伯特. 结构方程模型检验: 拟合指数与卡方准则 [J]. 心理学报, 2004(2): 186–194.
[25] Gierl M J, Lai H, Turner S R. Using automatic item generation to create multiple–choice test items: Automatic generation of test items [J]. Medical Education, 2012, 46(8): 757–765.
[26] Hernandez I, Nie W. The AI - IP: Minimizing the guesswork of personality scale item development through artificial intelligence [J]. Personnel Psychology, 2023, 76(4): 1011–1035.
[27] Gierl M J, Haladyna T M. Automatic item generation: Theory and practice [M]. Routledge, 2013.

https://doi.org/10.35534/pc.0603037

［28］Gierl M J，Lai H．Using Automatic Item Generation to Create Solutions and Rationales for Computerized Formative Testing［J］．Applied Psychological Measurement，2018，42（1）：42–57.

Exploration of the Development and Validation of the Reliability and Validity of the Chinese Emotional Stability Questionnaire based on ChatGPT

Gao Yaojie

Faculty of Psychology, Tianjin Normal University, Tianjin

Abstract: As emotional stability has received more and more attention from society and researchers, and in view of its important role in national health and several research fields (e.g., organizational management, mental health, education, physical health, decision-making behavior, etc.), good-quality emotional stability questionnaires have become particularly crucial. However, emotional stability as a personality trait is currently measured by the neuroticism dimension of the Big Five scales, which is not conducive to public understanding and dissemination and also exposes the fact that the localized measurement tools need to be enriched. Therefore, this study aims to develop a Chinese version of the Emotional Stability Questionnaire and, in line with the development trend of artificial intelligence, learns from the latest foreign paradigm of item development and explores the feasibility of using the latest version of ChatGPT for automatic item generation to develop a Chinese version of the Emotional Stability Questionnaire. The formal administration of this study included 68 items generated by the GPT retained after expert screening and a total of 42 items measuring neuroticism from four widely used Big Five personality scales (CBF-PI-B, Chinese version of the BFI-2, translated TIPI-10 and IPIP-BFAS), with a total of 479 valid samples obtained. The performance of the GPT-generated items was verified through the testing steps of item analysis, structural validity and Criterion-Related Validity calculation, reliability analysis, and comparisons, resulting in a good-quality Chinese version of the Emotional Stability Questionnaire.

Key words: ChatGPT; Questionnaire development; Emotional stability; Reliability; Validity