State-of-the-Art Natural Language Processing in der Entwicklung von Persönlichkeitsskalen:
Eine psychometrische Evaluation maschinen-generierter Items

Persönlichkeitsskalen werden im Personalbereich vielseitig eingesetzt, um Vorhersagen über individuelles und zwischenmenschliches Verhalten am Arbeitsplatz zu treffen. Die prädiktive Validität kann dabei durch Kontextualisierung erheblich gesteigert werden, wie vorige Studien beispielsweise für die Vorhersage der Arbeitsleistung (Bing et al., 2014; Hunthausen et al., 2003; Swift & Peterson, 2019) und Arbeitszufriedenheit (Swift & Peterson, 2019) gezeigt haben. Durch die Anpassung eines Items aus einer generischen Persönlichkeitsskala an den Kontext, für den Vorhersagen gemacht werden sollen, wird der Interpretationsspielraum für die Bedeutung des Items eingeschränkt. Das verringert Inkonsistenzen innerhalb der Bewertungen durch eine Person und für diesen Kontext irrelevante Variabilität zwischen Personen (Ajzen, 1987; Lievens et al., 2008; Paunonen & Ashton, 2001). Trotzdem werden kontextualisierte Persönlichkeitsskalen in der Praxis nicht durchgängig eingesetzt. Für den Arbeitskontext gibt es zwar unteranderem das Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung (Hossiep & Bräutigam, 2010), die Occupational Personality Questionnaires (Saville & Holdsworth Ltd., 1984), das Work Behavior Inventory (Page, 2009) und das Performance Perspectives Inventory (Abraham & Morrison, 2009) und damit größere Auswahl als für andere Kontexte - gegeben der vielfältigen Anwendung von Persönlichkeitsskalen im Personalbereich und der weitreichenden Entscheidungen die sie zum Beispiel bei der Personalauswahl beeinflussen, scheint das Angebot insbesondere an spezialisierteren Fragebögen jedoch beschränkt. Eine Hürde für die Verwendung flexibler, kontextualisierter Persönlichkeitsskalen ist der kostspielige Skalenentwicklungsprozess (Götz et al., 2023; Nunnally, 1978). Holtrop et al. (2014) führen das in ihrer Studie explizit als Limitation der kontextualisieren Skalen auf: “Designing a completely contextualized questionnaire took roughly 65 h” (p. 239). Entsprechend steckt großes Potential in der Automatisierung der Entwicklung kontextualisierter Skalen.

Neueste Entwicklungen im Bereich der Künstlichen Intelligenz, insbesondere der generativen KI und Natural Language Processing (NLP), könnten hierfür nützlich sein. Jüngste Studien im Bereich der Automatisierten Itemgenerierung (AIG) für Persönlichkeitsskalen haben bereits die Fähigkeit von NLP, konkret GPT-2 und GPT-3 aufgezeigt, generische Persönlichkeitsitems zu generieren, die psychometrische Eigenschaften erfüllen und voll funktionierende Skalen bilden (Davier, 2018; Götz et al., 2023; Hommel et al., 2022; Lee et al., 2023). Basierend auf dieser

Forschung werden in dieser Studie vier kontextualisierte Skalen aus maschinengenerierten Items erstellt und einer strengen psychometrischen Auswahl und Validierung unterzogen. Zum ersten Mal in der AIG-Forschung, konzentriert sich diese Studie auf die Entwicklung von kontextualisierten anstatt generischen Persönlichkeitsskalen mithilfe von NLP und wandelt so Erkenntnisse aus anderen Studien über die Vorteile von Kontextualisierung in einen handlungsorientierten Ansatz um, der den Zugang zu höherer prädiktiver Validität erleichtern kann. Zwanzig kontextualisierte Items wurden mit einem einzigen One-Shot-Prompt generiert und wurden nicht geändert oder weiter ausgewählt. Das Auslassen der Expertenauswahl nach der Generierung stellt einen Ansatz dar, der in der Forschung zur AIG bisher nicht verfolgt wurde und der einen nächsten Schritt in der Automatisierung der Entwicklung von Persönlichkeitsskalen darstellt.

Methode

Im ersten Schritt wurden Persönlichkeitsitems für die Eigenschaften Gewissenhaftigkeit und Ehrlichkeit-Bescheidenheit mithilfe eines GPT-3.5 basierten Prompts generiert. Um die Verlässlichkeit der maschinen-generierten Items umfassend zu untersuchen, wurden Items für Gewissenhaftigkeit und Ehrlichkeit-Bescheidenheit für zwei verschiedene Kontexte generiert: für den Arbeitsplatz und romantische Beziehungen.

Im zweiten Schritt wurden 335 Personen über Prolific rekrutiert. Den Teilnehmenden wurden kontextualisierte, maschinen-generierte Items sowie generische, menschen-generierte Items und Kriteriumsvariablen in einem within-subject design vorgelegt. Zur Erfassung der Persönlichkeitseigenschaften mit den generischen, menschen-generierten Skalen wurden die 12-Item Gewissenhaftigkeitsskala aus dem IPIP-NEO-60 (Maples-Keller et al., 2019) und die 10-Item Ehrlichkeit-Bescheidenheitsskala aus dem HEXACO-60 (Ashton & Lee, 2008) verwendet. Als Kriteriumsvariablen wurden Leistung, Zufriedenheit und Tugend jeweils für den Arbeits- und Beziehungskontext mit passenden Skalen abgefragt. Die Aufteilung auf zwei Wellen diente dazu, Übertragungseffekte zwischen Prädiktor- und Kriteriumsvariablen zwischen den einzelnen Kontexten zu vermeiden. Anschließend wurden mithilfe explorativer Faktorenanalyse in Anlehnung an die AIG-Studie von Götz et al. (2023) vier maschinen-generierte Skalen extrahiert (eine 4-Item Skala pro Trait-Kontext Kombination). Diese maschinen-generierten Skalen wurden dann einer umfassenden psychometrischen Evaluation unterzogen und mit den Ergebnissen der menschen-generierten Skalen verglichen. Konkret wurden Reliabilität (Cronbachs Alpha, McDonalds Omega), Invarianz gegenüber Alter und Geschlecht (Chi^2, CFI, TLI, RMSEA, SRMR), Konstruktvalidität (Chi^2, CFI, TLI, RMSEA,

SRMR, interne Korrelationen) und Diskrimination (Graded Response Model Parameter, Response Option Characteristic Kurven, Testinformationsfunktionen) sowie inkrementelle Validität (hierarchische Regression) untersucht.

Ergebnisse und Diskussion

Die Ergebnisse zeigen, dass die maschinen-generierten Skalen überwiegend gute psychometrische Eigenschaften und inkrementelle Validität über eine Reihe von Kriteriumsvariablen hinweg aufweisen. Nachdem zwei verschiedene Kontexte (Arbeit und romantische Beziehung) und eine breitere Reihe von externen Kriteriumsvariablen evaluiert wurden, lässt sich annehmen, dass mithilfe von NLP die prädiktive Validität von Persönlichkeitsmessungen durch Kontextualisierung verbessert werden und gleichzeitig der mit dem Entwicklungsprozess solcher Skalen verbundene Aufwand über verschiedene Bereiche hinweg reduziert werden kann.

Während die maschinen-generierten Skalen im Allgemeinen inkrementelle Validität für die meisten Kriteriumsvariablen aufweisen, variiert das Ausmaß der zusätzlich erklärten Varianz durch diese Skalen stark. Für die maschinen-generierten Skalen im Arbeitskontext fällt die Verbesserung der prädiktiven Validität gegenüber etablierten Skalen eher gering aus (zwischen 0.01 und 0.12 zusätzlich erklärte Varianz), während bei den maschinen-generierten Skalen für den Kontext romantischer Beziehungen die Verbesserung größer ist (zwischen 0.06 und 0.24 zusätzlich erklärte Varianz).

Es ist wichtig anzumerken, dass wir auch bei den menschen-generierten Skalen in bestimmten Aspekten Einschränkungen festgestellt haben. Zum Beispiel entsprach die Modellanpassung nicht den etablierten Standards. Dies deutet darauf hin, dass der vorgestellte Ansatz zur Automatisierung nicht nur eine Ergänzung ist, sondern ein wertvoller Schritt in Richtung einer verbesserten Persönlichkeitsmessung.

Die Automatisierung der Itemgenerierung und der Verzicht auf die darauffolgende Expertenselektion decken die ersten zwei Schritte der Skalenentwicklung ab. Die Ergebnisse ermutigen weiter zu untersuchen, ob auch nachfolgende Schritte wie die Befragung und anschließende Itemselektion mithilfe von NLP automatisiert werden können. Für ersteres könnte sich ein GPT-Prompt anbieten, der die maschinen-generierten Items aus der Perspektive verschiedener Personen beantwortet. Letztlich ist das Ziel, mithilfe von NLP eine vollständig einsatzbereite Persönlichkeitsskala zum interessierenden Konstrukt und Kontext generieren zu können, die etablierten Standards entspricht oder diese sogar übersteigt.

Literaturverzeichnis

Abraham, J., & Morrison, J. D. (2009). *Performance perspectives inventory: PPI technical manual, Version 8.*

Ajzen, I. (1987). Attitudes, Traits, and Actions: Dispositional Prediction of Behavior in Personality and Social Psychology. In *Advances in Experimental Social Psychology. Advances in Experimental Social Psychology Volume 20* (Vol. 20, pp. 1–63). Elsevier. https://doi.org/10.1016/S0065-2601(08)60411-6

Ashton, M. C., & Lee, K. (2008). The prediction of Honesty–Humility-related criteria by the HEXACO and Five-Factor Models of personality. *Journal of Research in Personality*, *42*(5), 1216–1228. https://doi.org/10.1016/j.jrp.2008.03.006

Bing, M. N., Davison, H. K., & Smothers, J. (2014). Item-level Frame-of-reference Effects in Personality Testing: An investigation of incremental validity in an organizational setting. *International Journal of Selection and Assessment*, *22*(2), 165–178. https://doi.org/10.1111/ijsa.12066

Davier, M. von (2018). Automated Item Generation with Recurrent Neural Networks. *Psychometrika*, *83*(4), 847–857. https://doi.org/10.1007/s11336-018-9608-y

Götz, F. M., Maertens, R., Loomba, S., & van der Linden, S. (2023). Let the algorithm speak: How to use neural networks for automatic item generation in psychological scale development. *Psychological Methods.* Advance online publication. https://doi.org/10.1037/met0000540

Holtrop, D., Born, M. P., Vries, A. de, & Vries, R. E. de (2014). A matter of context: A comparison of two types of contextualized personality measures. *Personality and Individual Differences*, *68*, 234–240. https://doi.org/10.1016/j.paid.2014.04.029

Hommel, B. E., Wollang, F.-J. M., Kotova, V., Zacher, H., & Schmukle, S. C. (2022). Transformer-Based Deep Neural Language Modeling for Construct-Specific Automatic Item Generation. *Psychometrika*, *87*(2), 749–772. https://doi.org/10.1007/s11336-021-09823-9

Hossiep, R., & Bräutigam, S. (2010). Personalauswahl und-entwicklung mit dem Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung (BIP). In W. Simon (Ed.), *GABAL professional training. Persönlichkeitsmodelle und Persönlichkeitstests: 15 Persönlichkeitsmodelle für Personalauswahl, Persönlichkeitsentwicklung, Training und Coaching* (2. Aufl., Vol. 15, pp. 136–158). Gabal-Verl.

Hunthausen, J. M., Truxillo, D. M., Bauer, T. N., & Hammer, L. B. (2003). A field study of frame-of-reference effects on personality test validity. *The Journal of Applied Psychology*, *88*(3), 545–551. https://doi.org/10.1037/0021-9010.88.3.545

Lee, P., Fyffe, S., Son, M., Jia, Z., & Yao, Z. (2023). A Paradigm Shift from "Human Writing" to "Machine Generation" in Personality Test Development: An Application of State-of-the-Art Natural Language Processing. *Journal of Business and Psychology*, *38*(1), 163–190. https://doi.org/10.1007/s10869-022-09864-6

Lievens, F., Corte, W. de, & Schollaert, E. (2008). A closer look at the frame-of-reference effect in personality scale scores and validity. *The Journal of Applied Psychology*, *93*(2), 268–279. https://doi.org/10.1037/0021-9010.93.2.268

Maples-Keller, J. L., Williamson, R. L., Sleep, C. E., Carter, N. T., Campbell, W. K., & Miller, J. D. (2019). Using Item Response Theory to Develop a 60-Item Representation of the NEO PI-R Using the International Personality Item Pool: Development of the IPIP-NEO-60. *Journal of Personality Assessment*, *101*(1), 4–15. https://doi.org/10.1080/00223891.2017.1381968

Nunnally, J. C. (1978). *Psychometric theory* (2. ed.). *McGraw-Hill series in psychology*. McGraw-Hill.

Page, R. C. (2009). *Work behavior inventory: Manual and user's guide.*

Paunonen, S. V., & Ashton, M. C. (2001). Big Five factors and facets and the prediction of behavior. *Journal of Personality and Social Psychology*, *81*(3), 524–539. https://doi.org/10.1037/0022-3514.81.3.524

Saville & Holdsworth Ltd. (1984). *Occupational Personality Questionnaires concept model manual & user's guide.*

Swift, V., & Peterson, J. B. (2019). Contextualization as a means to improve the predictive validity of personality models. *Personality and Individual Differences*, *144*(4), 153–163. https://doi.org/10.1016/j.paid.2019.03.007