Die stille Revolution in der Eignungsdiagnostik

Wie KI-Agenten die psychometrische Testung vollständig automatisieren

Dr. Walter Lieberei

23.08.2025

1. Die traditionelle Hürde der psychometrischen Testung

In einer sich rasant wandelnden Arbeitswelt ist die präzise Erfassung von Kompetenzen, Potenzialen und Persönlichkeitsmerkmalen entscheidender denn je. Psychometrische Testverfahren sind das Fundament einer objektiven und validen Eignungsdiagnostik. Doch ihre Entwicklung und Durchführung sind traditionell ein arbeits- und kostenintensiver Prozess. Die Konstruktion, Itementwicklung, Pilotierung, Auswertung, Normierung und das kontinuierliche Monitoring eines einzigen validen Tests erfordern oft interdisziplinäre Fachteams aus Psychologie, Statistik und IT sowie Entwicklungszyklen, die sich über Monate, wenn nicht sogar Jahre erstrecken.

Dieser immense Aufwand stellt für viele Unternehmen eine signifikante Hürde dar. Die Folge: Man greift auf standardisierte, oft teuer eingekaufte Tests zurück, die nicht immer passgenau auf die spezifischen Anforderungen des Unternehmens oder der zu besetzenden Stelle zugeschnitten sind. Die mangelnde Agilität dieses Prozesses steht im direkten Widerspruch zu den Anforderungen einer dynamischen Personal- und Organisationsentwicklung.

Doch wir stehen an der Schwelle zu einer Revolution. Fortschritte in den Bereichen der künstlichen Intelligenz (KI), insbesondere bei großen Sprachmodellen (LLMs) und autonomen Agenten-Architekturen, ermöglichen heute, was gestern noch als Zukunftsvision galt: die weitgehende Automatisierung des gesamten Testentwicklungs- und -durchführungsprozesses – von der ersten Literaturrecherche bis zur Auslieferung an den Endnutzer. Dieser Artikel beleuchtet diesen Paradigmenwechsel, erklärt die technologischen Bausteine und skizziert ein Vorgehensmodell für die voll automatisierte, KI-gesteuerte Testentwicklung.

2. Der klassische Weg: Ein Blick auf die traditionelle Testentwicklung

Um das Ausmaß der KI-gesteuerten Revolution zu verstehen, ist ein kurzer Blick auf den klassischen Prozess der Testkonstruktion unerlässlich (s. Tab. 1; vgl. Höft & Kersting, 2018). Dieser Prozess ist standardisiert und in klare Phasen unterteilt, um die wissenschaftliche Güte (Objektivität, Reliabilität und Validität) der Verfahren sicherzustellen.

Tab. 1: Typische Ablaufschritte bei der traditionellen Verfahrenskonstruktion

Schritt	Phase	Beschreibung
0	**Vorarbeiten**	Anforderungsanalyse und Festlegung der zu erfassenden Eignungsmerkmale. Entscheidung für eine bestimmte Testart (z. B. Leistungstest, Persönlichkeitstest, Simulation).

Schritt	Phase	Beschreibung
1	**Entwurf eines Szenarios**	Entwicklung von beruflichen Problemsituationen oder Aufgabenstellungen, die das relevante Verhalten provozieren sollen.
2	**Konstruktion Prototyp**	Ausgestaltung eines ersten Entwurfs mit allen Materialien, Instruktionen und Beobachtungs- bzw. Beurteilungsbögen.
3	**Erprobung Prototyp**	Durchführung von Testläufen mit “Testpersonen”, die der Zielgruppe ähneln, um Schwierigkeit, Realitätsnähe und Verständlichkeit zu prüfen.
4	**Revision & Neuerprobung**	Überarbeitung des Prototyps basierend auf dem Feedback und den Beobachtungen aus den Testläufen. Gegebenenfalls folgen weitere Erprobungsdurchläufe.
5	**Finalisierung**	Erstellung der finalen Testversion inklusive aller Handhabungshinweise, Auswertungsschlüssel und ggf. Parallelversionen zur Vermeidung von Lerneffekten.
6	**Normierung (optional)**	Erhebung von Daten einer großen, repräsentativen Stichprobe, um Vergleichswerte (Normen) zu erstellen, die eine Einordnung individueller Testergebnisse ermöglichen.
7	**Qualitätssicherung**	Kontinuierliche Überwachung der Testgütekriterien im laufenden Betrieb und gegebenenfalls Neukalibrierung der Normen nach einigen Jahren.

Dieser Prozess basiert auf zwei fundamentalen Theorien:

- **Klassische Testtheorie (KTT):** Ein pragmatischer Ansatz, bei dem angenommen wird, dass sich ein beobachteter Testwert aus einem “wahren” Wert und einem Messfehler zusammensetzt. Ziel ist es, den Messfehler durch sorgfältige Itemkonstruktion und statistische Analysen (z. B. Item-Schwierigkeit, Trennschärfe, Reliabilität) zu minimieren.
- **Probabilistische Testtheorie (PTT) / Item-Response-Theorie (IRT):** Ein anspruchsvolleres Modell, das die Wahrscheinlichkeit einer bestimmten Antwort in Abhängigkeit von der Fähigkeit einer Person und den Eigenschaften des Items (z. B. Schwierigkeit, Trennschärfe, Ratewahrscheinlichkeit) modelliert. IRT-basierte Tests ermöglichen z. B. computer-adaptives Testen, bei dem die Aufgabenschwierigkeit an das Leistungsniveau des Teilnehmers angepasst wird.

Beide Ansätze erfordern in der traditionellen Umsetzung einen hohen manuellen Aufwand, insbesondere bei der Recherche und der Formulierung sowie empirischen Erprobung der einzelnen Test-Items.

3. Der KI-Ansatz: Von der Teil- zur Vollautomatisierung

Die Idee, Technologie zur Unterstützung der Testentwicklung zu nutzen, ist nicht neu. Ein zentraler Meilenstein war und ist die **Automatisierte Item-Generierung (AIG)**, die sich in den letzten zwei Jahrzehnten als eigenes Forschungsfeld etabliert hat.

3.1 Automatisierte Itemgenerierung (AIG)

AIG bezeichnet den Prozess, bei dem Testaufgaben (Items) nicht mehr einzeln manuell geschrieben, sondern mithilfe von Computeralgorithmen aus vordefinierten Modellen oder Schemata erzeugt werden. Dies hat das Potenzial, die Anzahl der verfügbaren Items rapide zu erhöhen und den Prozess effizienter zu gestalten (Embretson & Kingston, 2018).

Man unterscheidet grundsätzlich zwei Ansätze (Circi, Hicks, & Sikali, 2023):

1. **Template-basierte AIG:** Hier werden sogenannte “Item-Modelle” erstellt. Ein Item-Modell ist eine Art Schablone, die die Struktur eines Items definiert (z. B. den Stamm einer Frage, die Art der Antwortoptionen) und Variablen enthält, die systematisch mit Inhalten gefüllt werden.
    - **Beispiel (vereinfacht):** Ein Item-Modell für eine mathematische Textaufgabe könnte Variablen für Namen, Objekte und Zahlen enthalten. Der Algorithmus befüllt diese Variablen aus Datenbanken und generiert so hunderte ähnlicher, aber nicht identischer Aufgaben. Studien zeigten, dass mit diesem Ansatz hohe Erfolgsraten in der qualitativen Begutachtung erzielt werden können, oft über 80-90% (Embretson & Kingston, 2018).
2. **Nicht-Template-basierte AIG:** Dieser Ansatz nutzt fortschrittliche Methoden des Natural Language Processing (NLP) und des maschinellen Lernens, um Items direkt aus unstrukturierten Textquellen (z. B. Fachartikeln, Lehrbüchern) zu generieren. Hierbei analysiert die KI den Text, identifiziert Schlüsselkonzepte und formuliert eigenständig Fragen und Distraktoren.

Während AIG bereits einen enormen Effizienzgewinn darstellt, bleibt der Prozess oft auf die Item-Erstellung beschränkt und erfordert weiterhin erhebliche menschliche Expertise zur Erstellung der Modelle und zur Steuerung des Gesamtprozesses. Der nächste logische Schritt ist die vollständige Automatisierung durch agentische KI.

4. Das Vorgehensmodell: Ein sechsstufiger Workflow für den KI-Testentwicklungs-Agenten

Ein vollautomatisierter Prozess kann durch einen übergeordneten Steuerungs-Agenten orchestriert werden, der für jede Phase spezialisierte Sub-Agenten aktiviert. Das folgende Modell beschreibt einen vollständigen End-to-End-Workflow.

Phase 1: Bedarfsanalyse und Konzeption

- **Trigger:** Ein HR-Business-Partner oder eine Führungskraft definiert ein zu untersuchendes Konstrukt, z. B. “emotionale Resilienz bei Führungskräften im Projektmanagement”.
- **Systematische Recherche (Research-Agent):** Der KI-Agent führt eine umfassende, systematische Recherche in wissenschaftlichen Datenbanken (z. B. Google Scholar, PsycINFO) und im Web durch. Er identifiziert bestehende Theorien, Modelle, empirische Studien und bereits validierte Testverfahren. Gleichzeitig

analysiert er die interne Marktnachfrage und potenzielle Anwendungsfälle auf der unternehmenseigenen Diagnose- und Lernplattform.

- **Synthese & Entscheidung:** Der Agent synthetisiert die Ergebnisse zu einem Bericht und bewertet auf Basis vordefinierter Kriterien (z. B. wissenschaftliche Fundierung, Messbarkeit, Marktlücke, ethische Vertretbarkeit), ob die Entwicklung eines neuen Tests gerechtfertigt ist. Der Output ist eine “Go/No-Go”-Entscheidung mit Begründung und einem Vorschlag für den Testtyp (z. B. Persönlichkeitstest basierend auf PTT/IRT) und eine vorläufige Struktur.

Phase 2: Testkonstruktion und Itementwicklung

- **Dimensionen-Definition:** Basierend auf der Recherche definiert der Agent die zu messenden Dimensionen (Subskalen) des Konstrukts (z. B. “Stressresistenz”, “Impulskontrolle”, “Optimismus”).
- **Automatisierte Item-Generierung (Item-Agent):** Der Agent nutzt ein großes Sprachmodell (LLM), um einen Pool an Items basierend auf den definierten Dimensionen und Spezifikationen zu generieren (z. B. 5-stufige Likert-Skala).
- **Item-Filterung und -Kuratierung:** Die generierten Items werden automatisiert auf sprachliche Qualität, Eindeutigkeit, Redundanz und potenzielle Verzerrungen (Bias) analysiert. Ein zweites, unabhängiges KI-Modell kann zur Prüfung der inhaltlichen Validität (Face Validity) herangezogen werden.

Phase 3: Pilotierung und Datenerhebung

- **Stichproben-Planung (Data-Agent):** Der Agent bestimmt die erforderliche Stichprobengröße für eine Pilotstudie und definiert die soziodemografischen Merkmale der Zielstichprobe.
- **Integration mit Datenerhebungs-Plattform:** Der Agent erstellt automatisiert eine neue Studie auf einer Plattform wie Prolific oder Amazon Mechanical Turk via API und übermittelt den Test, die Stichprobenkriterien und die Vergütung.
- **Automatisierte Überwachung:** Der Agent überwacht den Fortschritt der Datenerhebung kontinuierlich und sendet eine Benachrichtigung, sobald die Zieldatengröße erreicht ist.

Phase 4: Psychometrische Auswertung und Test-Revision

- **Datenabruf und -bereinigung:** Der Agent holt die Rohdaten via API ab und bereinigt den Datensatz automatisiert (z. B. Entfernung von unvollständigen Datensätzen).
- **Algorithmische Testanalyse (QA-Agent):** Der Agent führt eine vollautomatische psychometrische Analyse durch:
    - **Itemanalyse:** Berechnung von Schwierigkeit, Trennschärfe, etc.
    - **Faktorenanalyse:** Prüfung, ob die Items wie theoretisch angenommen auf die definierten Dimensionen laden.
    - **Reliabilitätsanalyse:** Berechnung der internen Konsistenz (z. B. Cronbachs Alpha).

- **Iterative Test-Optimierung:** Basierend auf den Analyseergebnissen werden Items mit schlechten Kennwerten identifiziert und eliminiert. Eine Entscheidungslogik prüft, ob die Gütekriterien nach der Eliminierung noch erfüllt sind. Falls nicht, kann der Agent entscheiden, zu Phase 2 zurückzukehren, um neue Items zu generieren.
- **Human-in-the-Loop:** An diesem Punkt wird ein auto-generierter Bericht mit den Ergebnissen und KPIs an einen menschlichen Experten (z. B. einen approbierten Psychologen) gesendet. Dieser gibt eine finale "Go/No-Go"-Entscheidung für die nächste Phase. Dies ist ein entscheidender Schritt zur Sicherung ethischer und rechtlicher Standards (z. B. DIN 33430).

Phase 5: Normierung und Finalisierung

- **Normierungsstichprobe:** Falls erforderlich, initiiert der Agent eine zweite, größere Datenerhebung mit einer repräsentativen Stichprobe, um Normwerte zu erheben.
- **Normwertberechnung:** Der Agent berechnet relevante Normwerte (z. B. Prozentränge, T-Werte) und erstellt die entsprechenden Normtabellen.
- **Feedback-Generator entwickeln:** Der Agent entwickelt Textbausteine für die automatisierte Interpretation der möglichen Testergebnisse und erstellt eine Vorlage für einen dynamischen PDF-Feedbackbogen.

Phase 6: Bereitstellung und Monitoring

- **Implementierung (Deployment-Agent):** Der finale Test wird über eine interne API in die Testbibliothek der Diagnose- und Lernplattform eingespielt.
- **Verfügbarmachung:** Der Test wird für die Zielnutzer zur eigenständigen Durchführung freigeschaltet.
- **Kontinuierliches Monitoring (Monitoring-Agent):** Der Agent sammelt fortlaufend anonymisierte Nutzerdaten, überprüft in regelmäßigen Abständen die psychometrischen Gütekriterien und kalibriert die Normen bei Bedarf neu (z. B. alle 1-3 Jahre), um deren Aktualität zu gewährleisten.

5. Anwendungsfall: Skill-Management in einem Unternehmen

Stellen wir uns ein Technologieunternehmen vor, das eine interne "SkillTech"-Plattform zur Diagnose und Entwicklung von Mitarbeiterkompetenzen nutzt.

- **Szenario:** Die Unternehmensleitung stellt fest, dass in agilen Projekten zunehmend Konflikte aufgrund mangelnder **"kollaborativer Problemlösefähigkeit"** auftreten. Eine entsprechende Weiterbildungsmaßnahme soll entwickelt werden, doch zunächst bedarf es eines validen Instruments zur Messung der Ausgangslage.
- **Automatisierter Workflow:**
    1. Ein Personalentwickler gibt das Konstrukt "kollaborative Problemlösefähigkeit im agilen Kontext" in das System ein.

2. Der **Research-Agent** durchforstet wissenschaftliche Paper und identifiziert die Kerndimensionen: kognitive (z. B. gemeinsame mentale Modelle), soziale (z. B. Konfliktlösung) und motivationale Aspekte. Er findet keine exakt passenden, frei verfügbaren Tests. **Entscheidung: Go.**
3. Der **Item-Agent** generiert basierend auf diesen Dimensionen 150 situative Urteilsfragen (Situational Judgement Test Items), in denen kurze Fallbeispiele aus dem agilen Alltag geschildert werden.
4. Der **Data-Agent** rekrutiert über eine interne Schnittstelle 500 Mitarbeiter aus verschiedenen IT-Projekten für eine anonymisierte Online-Pilotstudie.
5. Der **QA-Agent** analysiert die Daten. Die Faktorenanalyse bestätigt drei Dimensionen. 25 Items werden wegen schlechter Trennschärfe eliminiert. Die Reliabilität (Cronbachs Alpha) der Skalen liegt bei > 0.85. Der Agent erstellt einen Validierungsbericht.
6. Ein **menschlicher Gatekeeper** (leitender Psychologe des Unternehmens) erhält den Bericht, prüft die verbleibenden Items stichprobenartig auf inhaltliche Plausibilität und gibt die finale Freigabe.
7. Der **Deployment-Agent** kompiliert die finale Testversion, integriert sie in die SkillTech-Plattform und macht sie für Projektteams verfügbar.
8. Der **Monitoring-Agent** überwacht die eingehenden Daten und meldet nach 1000 Durchführungen, dass die Gütekriterien stabil sind.

- **Ergebnis:** Innerhalb von vier bis sechs Wochen hat das Unternehmen ein maßgeschneidertes, wissenschaftlich fundiertes diagnostisches Instrument entwickelt – ein Prozess, der traditionell über ein Jahr gedauert hätte.

6. Fazit: Chancen und Herausforderungen der vollautomatisierten Testentwicklung

Die voll automatisierte Entwicklung psychometrischer Tests ist keine Zukunftsvision mehr, sondern technisch und wirtschaftlich realisierbar. Die Vorteile sind immens, doch es gibt auch klare Herausforderungen und Risiken, die ein robustes Governance-Modell erfordern.

Stärken und Chancen:

- **Geschwindigkeit & Kosten:** Die Entwicklungszeit kann von 12-18 Monaten auf 4-6 Wochen reduziert werden. Die Kosteneinsparung wird auf 40-60% geschätzt, hauptsächlich durch die Reduktion manueller Expertenarbeit.
- **Skalierbarkeit & Individualisierung:** Es können schnell und einfach Tests für sehr spezifische Nischenanforderungen (z. B. für eine einzelne Schlüsselposition) entwickelt werden, für die sich eine traditionelle Entwicklung nie gelohnt hätte.
- **Objektivität & Konsistenz:** Algorithmische Prozesse können menschliche Inkonsistenzen in der Item-Entwicklung reduzieren.
- **Kontinuierliche Verbesserung:** Durch permanentes Monitoring können Tests dynamisch aktuell gehalten und Normen regelmäßig angepasst werden.

Schwächen und Risiken:

- **Ethische & rechtliche Absicherung:** Die größten Herausforderungen liegen in der ethischen Absicherung, Fairness-Kontrolle und regulatorischen Compliance.
  - **Bias:** LLMs können in ihren Trainingsdaten vorhandene gesellschaftliche Vorurteile (Bias) reproduzieren. Dies erfordert Fairness-Audits und Differential-Item-Functioning (DIF)-Analysen.
  - **Nachvollziehbarkeit:** "Black-Box"-LLMs können DIN-Normen (wie die DIN 33430) verletzen, die Nachvollziehbarkeit fordern. Der Einsatz von "Explainable AI"-Layern ist notwendig.
  - **Datenschutz (DSGVO):** Die Nutzung von Cloud-LLMs mit sensiblen personenbezogenen Daten ist heikel. EU-gehostete Modelle und eine strikte "Privacy-by-Design"-Architektur sind erforderlich.
- **Qualität der generierten Items:** Die Verlässlichkeit von LLM-generierten Items erfordert eine mehrschichtige Validierung (KI-Selbstkritik, regelbasierte Checks, Human-Review). Eine rein maschinelle Erstellung ohne menschliche Endkontrolle ist aktuell, insbesondere in hochsensiblen Kontexten, nicht zu verantworten.
- **Menschliche Expertise:** Der Prozess eliminiert den menschlichen Experten nicht, sondern verändert seine Rolle. Er wird vom Ersteller zum Überwacher, Kurator und strategischen Entscheider (Gatekeeper). Die Fachaufsicht durch einen qualifizierten Psychologen bleibt unerlässlich.

Zusammenfassend lässt sich sagen: Unternehmen, die KI-Agenten in ein robustes Governance-Modell einbetten und eine minimale, aber entscheidende menschliche Supervision beibehalten, können ihre eignungsdiagnostischen Prozesse revolutionieren. Sie verkürzen nicht nur Entwicklungszyklen drastisch, sondern sichern und verbessern gleichzeitig die Qualität ihrer Personalauswahl und -entwicklung. Der voll automatisierte KI-Testentwickler ist ein mächtiges Werkzeug – es liegt an uns, ihn weise und verantwortungsvoll einzusetzen.

Literaturverzeichnis

Blum, D., & Holling, H. (2018). Automatic Generation of Figural Analogies With the IMak Package. *Frontiers in Psychology*, *9*, 1286. https://doi.org/10.3389/fpsyg.2018.01286

Blum, D., Holling, H., Galibert, M. S., & Forthmann, B. (2016). Task difficulty prediction of figural analogies. *Intelligence*, *56*, 72–81. https://doi.org/10.1016/j.intell.2016.03.001

Circi, R., Hicks, J., & Sikali, E. (2023). Automatic item generation: foundations and machine learning-based approaches for assessments. *Frontiers in Education*, *8*, 858273. https://doi.org/10.3389/feduc.2023.858273

Embretson, S. E., & Kingston, N. M. (2018). Automatic Item Generation: A More Efficient Process for Developing Mathematics Achievement Items? *Journal of Educational Measurement*, *55*(1), 112–131. https://doi.org/10.1111/jedm.12166

Höft, S., & Kersting, M. (2018). Anforderungsprofil, Verhaltensbeobachtung und Verhaltensbeurteilung. In Diagnostik- und Testkuratorium (Hrsg.), Personalauswahl kompetent gestalten: Grundlagen und Praxis der Eignungsdiagnostik nach DIN 33430 (S. 27–63). Springer. https://doi.org/10.1007/978-3-662-53772-5

Shin, E. (2021). *Automated Item Generation by Combining the Non-template and Template-based Approaches to Generate Reading Inference Test Items* [Unveröffentlichte Doktorarbeit]. University of Alberta.