Die Persönlichkeit ist komplex – sie zu testen ebenso

Lassen sich Antworttendenzen und Manipulation in Persönlichkeitsfragebögen durch neue Itemformate vermeiden? Eine kurze Einführung für die praktische Beurteilung von Testformaten.

Von Luc Watrin und Andreas Frintrup

● Das Ziel von Personalauswahl ist es, auf Basis verlässlicher Informationen vor der Einstellung auf das Leistungsvermögen nach der Einstellung zu schließen. Dafür werden auch Persönlichkeitsfragebögen als Auswahlverfahren genutzt. Legitimiert wird dies durch eine Vielzahl empirischer Arbeiten, die einen Zusammenhang von Persönlichkeitseigenschaften und Leistungsvermögen nachweisen. Doch wie lässt sich Persönlichkeit im Fragebogen am zuverlässigsten erfassen, insbesondere im Auswahlsetting?

Klassische Itemformate von Persönlichkeitstests

Persönlichkeitsfragebögen sind Selbstbeschreibungen einer Person in Bezug auf sich selbst. Bei der Bearbeitung werden die Teilnehmer deshalb gebeten, über die Ausprägung bestimmter Eigenschaften und Einstellungen bei sich selbst zu berichten. Dazu werden in der Psychologie üblicherweise sogenannte „Skalen" zur Selbsteinschätzung vorgegeben, auf denen die Zustimmung oder Ablehnung zu einer bestimmten Aussage angegeben werden kann. Hierfür stehen unterschiedliche Formate zur Verfügung, mit denen sich Teilnehmer differenziert einschätzen können. Skalen dürfen dabei weder zu undifferenziert noch zu differenziert sein. Infolge intensiver Forschung haben sich deshalb ungerade Skalenlängen mit 5, 7 oder 9 möglichen Markierungen und der Möglichkeit einer Antwort auch in der Skalenmitte durchgesetzt. Man nennt diese Skalen auch Likert-Skalen, benannt nach ihrem Entwickler Rensis Likert.

Der große Vorteil dieses Skalentyps ist die einfache und intuitive Handhabung, eine leichte Umsetzbarkeit sowohl auf Papier als auch in elektronischen Medien, eine langjährige Bewährung in Psychologie sowie Sozial-, Markt- und Meinungsforschung und die Gewöhnung der Testteilnehmer an dieses Format. Solange bei der Messung und ihrem Ergebnis nichts auf dem Spiel steht, funktionieren diese Skalen auch gut. Beispiele hierfür sind Produktbewertungen im Internet, die Einschätzungen von eigenen Kompetenzen für Trainings- und Personalentwicklungsprogramme sowie generell alle anonymen Erhebungen. Kritische Stimmen bemängeln jedoch die Anfälligkeit der Skalen für Messungenauigkeiten, verursacht durch persönliche Antworttendenzen und den Effekt „sozial erwünschten Antwortverhaltens".

Antworttendenzen bezeichnen die generelle Neigung einer Person, auf Skalen beispielsweise eher in Extremen zu antworten, Antwortkategorien um den Mittelpunkt der Skala zu bevorzugen oder Aussagen grundsätzlich, losgelöst vom Inhalt, zuzustimmen. Da diese Tendenzen einen Einfluss auf die Bewertung der Aussagen haben, der unabhängig von der eigentlichen Persönlichkeitsausprägung der Person ist, können sie die Persönlichkeitsmessung und insbesondere den Vergleich über verschiedene Personen hinweg beeinträchtigen.

Personaler müssen keine Experten in der Testtheorie sein. Aber einige spezifische Details sollten sie kennen, um valide Tests einsetzen zu können.

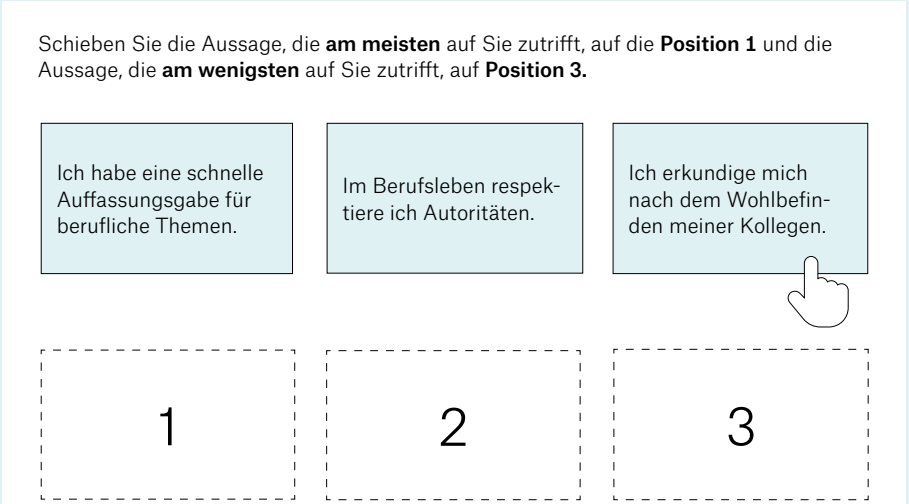
Hängt von dem Ergebnis eine für die Teilnehmer wichtige Bewertung oder sogar Entscheidung ab, laden Skalen außerdem zu einer „nützlichen" Veränderung des Messergebnisses ein, Bewerber stellen also Vermutungen über zweckdienliche Ausprägungen eines Merkmals an und antworten „sozial erwünscht". Gerade in eignungsdiagnostischen Situationen, wenn Bewerber also eine bestimmte Stelle gerne besetzen würden, wirken verschiedene Prozesse, die von einer eher unbewussten Bessereinschätzung bis zu Übertreibungen oder wirklicher Manipulation durch Falschbeschreibungen reichen können. Während die positive Selbstdarstellung vermeintlich sehr im Interesse der Bewerber liegt, ist eine falsche Selbstbeschreibung für Unternehmen und auch andere Bewerber (diejenigen, die ehrlich antworten) schädlich. Auch für die betroffenen Bewerber selbst ist die aktive Falschdarstellung langfristig eher Schaden als Nutzen – sie mögen zwar so die angestrebte Vakanz besetzen, müssen dann aber auf dem Job einlösen, was sie über sich selbst berichtet haben.

Die personalpsychologische Forschung hat viele Mühen investiert, um Effekte der Antworttendenzen und des sozial erwünschten Antwortverhaltens zu reduzieren und zu objektiveren, im engeren Sinne zu „ehrlicheren", Selbsteinschätzungen von Personen zu gelangen. Denn auch wenn sich Persönlichkeitstests mit Likert-Skalen trotz ihrer vermeintlichen Schwächen vielfach darin bewährt haben, beruflichen Erfolg vorherzusagen, weisen die Stimmen der Kritiker durchaus auf mögliches Verbesserungspotenzial hin.

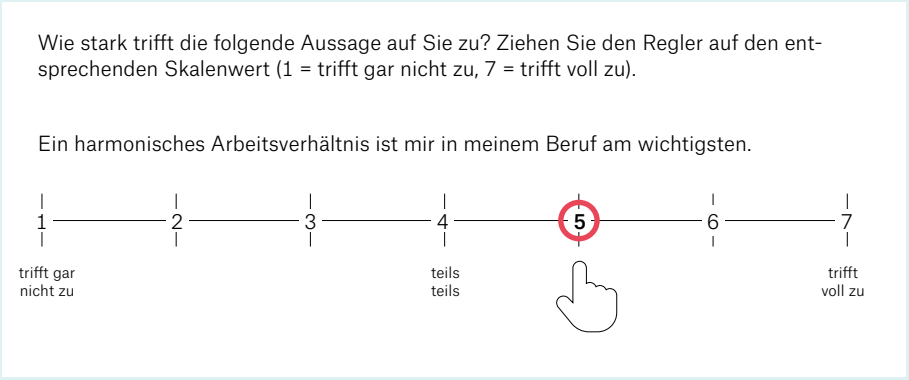
Das Forced-Choice-Format als Alternative zur Likert-Skala

Zur Lösung der beschriebenen Probleme wurde das Forced-Choice-(FC-)Format als alternatives Antwortformat für Persönlichkeitsfragebögen vorgeschlagen. Statt immer nur eine Aussage mit einer Antwortskala zu präsentieren, werden im FC-Format zwei oder mehr Aussagen gleichzeitig aufgezeigt. Bei zwei Aussagen müssen Testteilnehmer diejenige auswählen, die am meisten auf die eigenen Einstellungen und Verhaltensweisen zutrifft. Bei mehr als zwei Aussagen pro Block muss angegeben werden, welche Aussage am meisten und welche am wenigsten auf die eigene Persönlichkeit zu-

Item eines Persönlichkeitstests im Forced-Choice-Format



Konventionelle Likert-Skala mit sieben Antwortstufen



Quelle: HR Diagnostics AG

trifft beziehungsweise müssen diese in eine individuell passende Rangreihe gebracht werden. Ein Item eines FC-Persönlichkeitsfragebogens kann so aussehen, wie in der Grafik oben dargestellt.

Das FC-Format weist gegenüber Likert-Skalen eine Reihe vorteilhafter Eigenschaften auf. Die Auflösung der Skala zugunsten von Paarvergleichen oder Rangreihen löst die Problematik der beschriebenen Antworttendenzen vollständig auf.

Das Format fordert die Kandidaten außerdem dazu auf, sich intensiver mit den gestellten Aussagen auseinanderzusetzen, die ihre Persönlichkeit betreffen. So müssen Bewerber im Beispiel reflektieren, ob es ihnen im Beruf wichtiger ist, sich mit Kollegen gut zu verstehen oder ob sie sich stärker, als sie sich für das Wohlergehen von Kollegen interessieren, für eine Person mit hoher Auffassungsgabe und Leistungsfähigkeit halten. Diese Selbstreflexion kann für Kandidaten ebenso wertvoll sein wie für die Persönlichkeitsmessung an sich.

Wie das Beispiel verdeutlicht, ist es auch im FC-Format möglich, unterschiedliche Merkmale innerhalb einer Aufgabe zu testen, indem die Aussagen miteinander kombiniert werden. Ein zentraler Vorteil des FC-Formats: Es ist nicht möglich, sich frei von jeglichen Makeln zu präsentieren, wie es zum Beispiel manche Bewerbungsratgeber empfehlen. Ein Beispiel: Eine unattraktive Aussage zur Eigenschaft „Emotionale Stabilität“ („Ich fühle mich schnell bedroht.“) und eine Aussage zur „Verträglichkeit“ („Ich interessiere mich nicht für die Probleme anderer.“) werden in einem Block gemeinsam präsentiert. Statt in einer Bewerbungssituation beide Aussagen auf einer Likert-Skala niedrig bewerten zu können, müssen Bewerber nun eine Auswahl treffen, welche Verhaltensweise auf sie eher zutrifft – eine Information, die für Personalverantwortliche von großem Interesse ist. Auf diese Weise kann überzogene Selbstdarstellung reduziert werden, da Bewerber damit konfrontiert werden, mehrere Aussagen gegeneinander abzuwägen und eine konkrete Rangreihe zu bilden.

Anstrengung und Akzeptanz aus Bewerbersicht

Doch auch das FC-Format ist nicht frei von Kritik. So stellt sich in der Praxis die Frage, ob es für Bewerber unverhältnismäßig anstrengend ist, sich in einem Testverfahren derart intensiv mit mehreren persönlichkeitsbezogenen Aussagen gleichzeitig auseinandersetzen zu müssen. Sicherlich ist es einfacher, einzelne Aussagen zu bewerten, als mehrere Aussagen gegeneinander abzuwägen. Gleichzeitig können Bewerber beim FC-Format keine unattraktiven Aussagen vermeiden, indem sie sie schlicht niedrig bewerten – das nährt die Sorge von Unternehmen, dass Bewerber den Test nicht akzeptieren und das Verfahren abbrechen, bevor Recruiter die Gelegenheit bekommen, sie kennenzulernen.

Um die tatsächliche Akzeptanz oder Ablehnung seitens der Bewerber zu messen, haben wir in einer Studie dieselben Aussagen eines Fragebogens sowohl im FC-Format als auch mit Likert-Skalen präsentiert und die Testteilnehmer gebeten, ihre Einschätzung im Sinne der Akzeptanz für das jeweilige Format dazu

abzugeben. Das Ergebnis: Die Unterschiede in der Akzeptanz waren unerheblich. Dieses Ergebnis wird auch im Realeinsatz bei Kunden bestätigt – über 99 Prozent der eingeladenen Kandidaten, die den Test beginnen, absolvieren ihn auch vollständig bis zum Ende, die Abbruchquote liegt also unter 1 Prozent.

Bezogen auf die kognitive Anstrengung der Testteilnehmer herrscht in Forschung und Praxis Konsens, dass bis zu drei gleichzeitig zu bewertende Aussagen innerhalb einer Aufgabe zumutbar sind und damit valide und reliable Ergebnisse erzielt werden.

Ergebnisinterpretation: Der Schein kann trügen

Ein zentraler Kritikpunkt bezüglich FC-Tests besteht in ihrer Auswertung und, in der Folge, der Gültigkeit ihrer Ergebnisse. Deshalb lohnt es sich hier genau hinzusehen. Während das Format keineswegs neu ist und in einer Vielzahl von Tests Anwendung findet, können die Ergebnisse erst seit statistischen Entwicklungen der letzten Jahre überhaupt für den direkten Vergleich unterschiedlicher Personen genutzt werden – die zentrale Anforderung eines Tests in der Personalauswahl, denn schließlich sollen Eignungsunterschiede zwischen verschiedenen Personen festgestellt werden.

Über diese gravierende Schwäche sehen viele Testanbieter gerne hinweg und werten das Format analog zu konventionellen Likert-Skalen über eine irgendwie geartete Form der Summenbildung aus, auf deren Basis jedoch keine interindividuellen Vergleiche möglich sind. Summiert man die Punkte eines FC-Tests auf, ergibt sich für alle Kandidaten derselbe Gesamtwert – eine Unterscheidung zwischen Kandidaten ist folglich nicht möglich, es kann nur die Ausprägung unterschiedlicher Persönlichkeitseigenschaften innerhalb einer Person betrachtet werden.

Zusätzlich unterschlägt diese Herangehensweise, dass die Entscheidungen der Kandidaten hinsichtlich der einzelnen Aussagen innerhalb eines Blocks keineswegs unabhängig sind, wie dies bei konventionellen Skalen der Fall ist, sondern die gleichzeitig dargebotenen Aussagen einen substanziellen Einfluss aufeinander haben.

Wir setzen deshalb bei der Auswertung auf das von Brown und Maydeu-Olivares entwickelte „Thurstonian Item Response Theory (IRT) Model“. Statistisch gesehen handelt es sich um ein IRT-Modell, das in einem Strukturgleichungsmodell geschätzt wird und dessen theoretische Grundlage auf L. L. Thurstone, einen zentralen Wegbereiter der empirischen Psychologie, zurückgeht. Statt Antworten einzeln zu betrachten, fließen in die Auswertung die Ergebnisse der Paarvergleiche zwischen den Aussagen ein. In einer Reihe empirischer Untersuchungen, durchgeführt durch unabhängige Forscherteams und uns selbst, konnte aufgezeigt werden, dass die mit dem FC-Format vormals assoziierten Einschränkungen mit dem Thurstonian IRT Model behoben werden konnten. Ein Einsatz in der Personalauswahl ist damit möglich.

LUC WATRIN arbeitet im Bereich Forschung und Entwicklung bei der HR Diagnostics AG. Er ist Experte für Testentwicklung und -Evaluation.

ANDREAS FRINTRUP führt als CEO die HR Diagnostics AG. Er ist Testautor und Experte für die Gestaltung personaldiagnostischer Prozesse in Personalauswahl und Personalentwicklung.

Während die Entwicklung und Auswertung von FC-Tests erheblich höhere Anforderungen an das Know-how von Testentwicklern stellt, bleibt die Ergebnisinterpretation aufseiten der Anwender genauso einfach wie bei konventionellen Fragebögen. Ergebnisse werden ebenfalls in Form von standardisierten Z-Werten oder Prozenträngen ausgegeben und können somit einfach zu direkten normativen Vergleichen verwendet oder mit anderen Testergebnissen zu einem Gesamtwert verrechnet werden.

Klassische Antwortskala versus Forced-Choice-Format

Welches Antwortformat ist also insgesamt besser? Zunächst einmal muss man festhalten: Sowohl die klassische Likert-Skala als auch das Forced-Coice-Format genügen den wissenschaftlichen Ansprüchen der Eignungsdiagnostik. Sie erzielen beide fundierte Ergebnisse, auf deren Basis richtige Personalentscheidungen getroffen werden können – vorausgesetzt, das FC-Format wird adäquat ausgewertet. Es gilt also für Anwender, die jeweiligen Vor- und Nachteile beider Formate individuell abzuwägen.

Eine kurze Entscheidungshilfe für die Praxis können wir noch an die Hand geben: Likert-Skalen sind der etablierte Standard und einfach zu bearbeiten, Bewerber sind an das Format gewöhnt. Unterschiedliche Skalen lassen sich einfach kombinieren und eine individuelle Normierung an der eigenen Stichprobe ist auch bei geringem Bewerberaufkommen möglich. FC-Tests erfordern für individuelle Normierungen größere Fallzahlen ($N > 400$, davor kann aber auf umfangreiche bestehende Normdatenbanken zurückgegriffen werden) und das derzeitige Angebot an verfügbaren Persönlichkeitsskalen ist geringer. Das Potenzial, soziale Erwünschtheit und Antwortverzerrungen zu reduzieren, liegt jedoch auf der Hand.

Insofern ein letzter Tipp für die Entscheidung: Eine Mischung beider Formate ist ebenfalls möglich, um die Abwechslung bei der Testbearbeitung zu erhöhen. So können Sie sich die Vorteile beider Formate für Ihre Personalauswahl zunutze machen. ■