Intelligenztests – eine Bauanleitung

Fähigkeiten auf verschiedenen Ebenen

Wie entsteht ein Intelligenztest? Welche Voraussetzungen muss ein guter Test erfüllen? Dieser Artikel ist ein Bericht direkt vom Puls der Forschung, denn die Autorin entwickelt im Rahmen ihrer Dissertation gerade einen eigenen IQ-Test.

Ich freue mich über eure Anregungen und Ideen zu dieser Reihe. Mailt mir an MERF@mensa.de!

Stellen wir uns Folgendes vor: Die Hochbegabungsforscherin T. G. B. will Intelligenz bei Grundschulkindern erfassen, findet aber kein Verfahren, das alle ihr wichtigen Aspekte abdeckt. Sie beschließt also, selbst einen Intelligenztest zu entwickeln. Nichts leichter als das!

Das ist natürlich schamlos übertrieben – die Konstruktion eines IQ-Tests ist aufwändig. Im psychologischen Sinne ist ein Test ein diagnostisches Verfahren, das bestimmte Aspekte menschlichen Erlebens und Verhaltens so erfasst, dass die Messung bestimmten Gütekriterien genügt. In diesem Artikel steht die Diagnostik der Intelligenz im Vordergrund.

Theoretisch …

Ein guter Test basiert auf einem theoretischen Modell, das inhaltlich zur Zielsetzung passt, und die Auswahl dazu ist groß! In der Forschung weitgehend akzeptiert ist das integrative Modell von Carroll (1993). Seiner Ansicht nach kommt es darauf an, wie sehr man bei der Betrachtung der Intelligenz ins Detail geht; folglich beinhaltet sein Modell mehrere Ebenen. Auf der ersten Ebene *Stratum I* finden sich eng umschriebene Fähigkeiten. *Stratum II* beinhaltet breiter gefasste Fähigkeiten wie Gedächtnis, Wissen und Verarbeitungsgeschwindigkeit, *Stratum III* die allgemeine intellektuelle Fähigkeit eines Menschen. In dieses Modell lassen sich quasi alle aktuellen Intelligenzkonzeptionen eingliedern.

Vielseitige Fragen

In der Praxis ist jedoch wichtig, welche Facetten dieses reichhaltigen Konstrukts erfasst werden sollen und warum: Genügt ein Teilaspekt, etwa figurale Analogien oder Matrizen, um die Intelligenz einer Person abzuschätzen? Möchte man den Erfolg eines Kindes in einer Fördermaßnahme vorhersagen, wozu sprachliche und numerische Aufgaben benötigt werden? Oder geht es um ein differenziertes und facettenreiches Abbild der Intelligenzstruktur wie beim Mensa-Aufnahmetest? Die Entscheidung für ein Modell hängt von der Antwort auf diese Fragen ab. Ein zweiter Aspekt betrifft die Testsituation selbst: Einzel- oder Gruppentest? Auf dieser Basis beginnt nun die Literaturrecherche, um herauszufinden, welche Testverfahren es überhaupt gibt, welche davon für die intendierte Altersstufe geeignet sind und welche Untertests zum gewählten Konzept passen.

Hat man geeignete Aufgabentypen gefunden, muss man diese entsprechend verändern oder auf der ihnen zugrunde liegenden Idee aufbauend eigene Aufgaben konstruieren, denn auch Testverfahren unterliegen dem Urheberrecht. Ab und zu stößt man auf das Problem, dass für den gewünschten Bereich noch keine brauchbaren Aufgaben existieren, sodass diese – beispielsweise auf der Basis von Theorien kognitiver Entwicklung – neu entwickelt werden müssen.

In der Regel muss man deutlich mehr Aufgaben konstruieren als der Test später beinhalten wird, denn leider bewähren sich in der Praxis nicht alle davon. Deshalb sollte man Pilotierungsphasen einplanen, in denen die Verständlichkeit der Instruktionen und der zeitliche Aufwand für die einzelnen Aufgaben überprüft werden. Schließlich dienen diese Ergebnisse dazu, geeignete Aufgaben für die Endfassung auszuwählen.

„Ein Fieberthermometer misst noch lange keinen IQ von 37, auch wenn bei wiederholter Messung vermutlich jeder dieses Ergebnis abläse."

Guter Test, schlechter Test?

Was macht einen guten Test aus? Gemäß der klassischen Testtheorie sollte ein Test objektiv, reliabel und valide sein.

Objektivität ist dadurch charakterisiert, dass Durchführung, Auswertung und Interpretation standardisiert sind, etwa durch explizit formulierte Instruktionen, klar definierte korrekte und falsche Lösungen, die sich mit Hilfe von Schablonen oder Computerprogrammen ermitteln lassen, und Hinweise zur Deutung der Ergebnisse. Egal, wer also den Test durchführt, auswertet oder interpretiert: Das Resultat sollte dasselbe sein.

Die *Reliabilität* bezeichnet die Zuverlässigkeit des Verfahrens: Nimmt man an, dass das zu erfassende Konstrukt stabil ist, sollte bei einer erneuten Testung ein ähnliches Ergebnis herauskommen. Ein zweiter Aspekt der Reliabilität ist die *interne Konsistenz* des Verfahrens: Alle Aufgaben, die sich unter dasselbe Konstrukt (etwa sprachliche Begabung) subsumieren lassen, sollten untereinander und mit dem Gesamtwert dieser Skala statistisch zusammenhängen.

Das dritte Hauptgütekriterium ist die *Validität* des Verfahrens, also die Frage, ob der Test misst, was er zu messen vorgibt. Ein Intelligenztest sollte mit ähnlichen Konstrukten zusammenhängen, aber nicht durch andere Konstrukte, etwa Prüfungsängstlichkeit, beeinflusst sein.

Da die drei Kriterien aufeinander aufbauen, muss ein objektives und reliables Verfahren nicht unbedingt valide sein: Ein Fieberthermometer misst noch lange keinen IQ von 37, auch wenn bei wiederholter Messung vermutlich jeder dieses Ergebnis abläse.

Ein Nebengütekriterium ist die *Normierung*. Jeder Test sollte auf einer ausreichend großen Stichprobe basieren, sodass sich individuelle Werte mit anderen Personen des gleichen Alters, Geschlechts oder Bildungsgrads vergleichen lassen. Der sogenannte *Flynn-Effekt* hat in den vergangenen Jahrzehnten dazu geführt, dass Menschen immer höhere Intelligenzwerte erzielen, ohne dass sie tatsächlich intelligenter geworden wären. Zugenommen hat jedoch die Menge an Informationen, die wir täglich verarbeiten müssen, sowie die Erfahrung mit Tests, die heute viel verbreiteter sind als noch vor 50 Jahren. Die „Intelligenzzunahme" bewegt sich im Bereich von etwa 10 Punkten pro Generation – eine ganze Menge, wenn man bedenkt, dass 15 Punkte bereits eine komplette Standardabweichung bedeu-

„Für einen Intelligenztest, der im Bereich der Hochbegabung differenzieren soll, dürfen die einzelnen Aufgaben (Items) nicht zu leicht sein."

ten! Entsprechend wichtig ist es, die Daten der Vergleichsstichprobe kontinuierlich zu aktualisieren, um die tatsächliche Intelligenz einer Person nicht zu überschätzen. Aktuelle Befunde deuten darauf hin, dass der Effekt inzwischen rückläufig ist.

Die *Fairness* schließlich besagt, dass ein Test keine Personengruppe (Geschlecht, Ethnie, Altersgruppe) systematisch benachteiligen darf.

Items: Schwere Kost

Für einen Intelligenztest, der im Bereich der Hochbegabung differenzieren soll, dürfen die einzelnen Aufgaben (Items) nicht zu leicht sein. Die *Schwierigkeit* ermittelt man über den prozentualen Anteil der Personen, die eine bestimmte Aufgabe gelöst haben.

Mindestens ebenso wichtig ist die *Trennschärfe* der Items, also die Frage, inwieweit jede einzelne Aufgabe zwischen Personen mit hohen und niedrigen Gesamtwerten differenziert. Erfasst wird diese über die *Korrelation* (ein statistisches Zusammenhangsmaß) zwischen dem einzelnen Item und dem Gesamtwert (ohne das jeweilige Item). Löst jemand ein trennscharfes Item, kann man daraus eher auf eine hohe Gesamtpunktzahl schließen als aus der Lösung eines weniger trennscharfen Items.

Voraussetzung für akzeptable Schwierigkeits- und Trennschärfewerte ist eine ausreichende *Streuung* beziehungsweise Variabilität, dass also nicht alle Personen dieselbe Lösung ankreuzen. Für einen normalen Intelligenztest sind Lösungswahrscheinlichkeiten im mittleren Bereich (übliche Grenzwerte liegen zwischen 20 und 80 Prozent) bei hoher Trennschärfe günstig; bei Tests, die eher am oberen Ende differenzieren sollen, liegt die Schwierigkeit entsprechend höher.

Sind nach der Pilotierungsphase die „guten" Items zu einem Test zusammengestellt, geht es an die Erhebung einer hinreichend großen Stichprobe; mehrere tausend Personen (beim Mensatest über 20 000) sind keine Seltenheit. Dann folgt die Auswertung: Haben sich die Items bewährt? Kommen auch die Subskalen (etwa rechnerisches Schlussfolgern) heraus, die man theoretisch konzipiert hatte? Hängen die ermittelten Werte mit Außenkriterien (etwa anderen Intelligenztests oder Schulleistungen) zusammen? Spannend ist die Entwicklung eines solchen Tests auf jeden Fall!

Tanja Gabriele Baudson

Über die Autorin

Tanja Gabriele Baudson ist Diplompsychologin, Romanistin und Tauchlehrerin. Sie arbeitet als Begabungsforscherin am Lehrstuhl für Hochbegabtenforschung und -förderung der Universität Trier.

Literatur:

- Carroll, J. B. (1993). Human Cognitive Abilities. Cambridge: Cambridge University Press.
- Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. Psychological Bulletin, 101, 171–191.