IQ-Interpretation: Was bedeuten die Grundraten?

oder: Wieso die 23-Punkte-Regel überflüssig ist

Autorin: Dipl.-Psych. Dr. Alexandra Lenhard (Dezember 2019)

In meinen Intelligenzseminaren merke ich immer wieder, dass viele Anwender*I*nnen große Schwierigkeiten haben, die Bedeutung von Diskrepanzen zwischen verschiedenen Skalen innerhalb einer Testbatterie (oder unter Umständen auch zwischen Testbatterien) richtig einzuordnen. Manchmal passiert es auch, dass ich langwierig über die Bedeutung von Grundraten in der WISC referiere, nur um anschließend von den Teilnehmern gesagt zu bekommen, man würde die 23-Punkte-Regel anwenden, man habe das schließlich so gelernt. Von einigen wird angeführt, diese Regel stünde doch sogar in den Testmanualen von WISC-V oder KABC-II – was aber nicht zutrifft. Neulich kam mir sogar zu Ohren, die Regel sei „Allgemeinwissen" eines jeglichen Testleiters oder einer jeglichen Testleiterin.

Die 23-Punkte-Regel

Für diejenigen Testanwender*I*nnen, welche die 23-Punkte-Regel nicht kennen, möchte ich sie an dieser Stelle kurz erklären: Sie besagt, dass die Aussagekraft eines Gesamtindex für die allgemeine Intelligenz (also z. B. des G-IQ in der WISC-V) stark eingeschränkt ist, sobald die Diskrepanz zwischen der höchsten und niedrigsten Indexskala (also z. B. zwischen FS und VG) 23 IQ-Punkte (d. h. 1,5 *SD*) oder mehr beträgt.

Nun ist es nicht so, dass diese Regel gänzlich in den Bereich der „Fake-News" zu verweisen ist. Tatsächlich findet man Sie sogar von den Autoren der K-ABC höchstpersönlich in den „Essentials of KABC-II Assessment" niedergeschrieben (A. S. Kaufman, Lichtenberger, Fletcher-Janzen & Kaufman, 2005, S. 86). Die Autoren merken allerdings an, die Regel sei „straightforward and easy to remember". Im Klartext bedeutet dies, dass es sich hierbei um eine stark vereinfachende Pi-mal-Daumen-Regel handelt, die man anwenden kann, wenn man zusätzlich zu den Skalenwerten nur seinen Kopf (inkl. der beschränkten Gedächtnis- und Rechenressourcen), nicht aber den vollständig ausgefüllten Auswertungsbogen der Intelligenztestung zur Verfügung hat. Da meiner Erfahrung nach die meisten Testanwender*I*nnen mittlerweile zum Glück auf computerbasierte Auswertung zurückgreifen, stehen für die Wechsler-Skalen und die KABC-II jedoch auch ohne den Rückgriff auf zusätzliche Regeln oder aufwändige Berechnungen wesentlich genauere Zahlenwerte zur Verfügung, auf die bei der Bewertung der Aussagekraft des IQ-Wertes zurückgegriffen werden kann und sollte. Die Anwendung der 23-Punkte-Regel wird damit im Prinzip überflüssig.

Manche Anwender*I*nnen haben jedoch vermutlich Schwierigkeiten damit, die entsprechenden Zahlenwerte auf dem Auswertungsbogen zu finden und richtig einzuordnen. Ich möchte deshalb an dieser Stelle versuchen, die einzelnen Schritte bei der Analyse von Skalendifferenzen möglichst prägnant zu erklären und mit kurzen Beispielen zu veranschaulichen.

Unterschiedliche Aspekte von Skalendifferenzen

Die Bedeutung einer Differenz zwischen zwei Skalen (oder zwischen einer Skala und dem Mittel aus allen Skalen) wird üblicherweise unter zwei Gesichtspunkten betrachtet. Der erste betrifft die Messsicherheit, also die Reliabilität der Skalen. Die zugrundeliegende Idee ist diejenige, dass eine Differenz zwischen zwei Messwerten nur dann interpretiert werden sollte, wenn die Wahrscheinlichkeit, dass sie ausschließlich durch Messfehler zustande kam (d.h., dass gar kein Unterschied der wahren Werte vorliegt), gering ist. In der Statistik bedeutet „geringe Wahrscheinlichkeit" in der Regel eine Auftretenshäufigkeit unter 5%. In der Individualdiagnostik verortet man diese Grenze (d. h., das Signifikanzniveau) anstatt bei 5% häufig auch bei 10% (vgl. Huber, 1973). Damit minimiert man das Risiko, auffällige Diskrepanzen fälschlicherweise auf Messfehler zurückzuführen und deshalb zu ignorieren.

Der Reliabilitätsaspekt

In einem ersten Schritt wird also bei der Analyse von Skalendifferenzen geprüft, ob die Differenz zwischen zwei Skalen (oder einer Skala und dem Mittel aus allen Skalen) über der sogenannten „kritischen Differenz" liegt. Letztere wird auf der Basis der Reliabilität der einzelnen Skalen und des gewählten Signifikanzniveaus berechnet (siehe z. B. Amelang & Schmidt-Atzert, 2006, S. 53) und in den Handbüchern spezifiziert bzw. von der Auswertungssoftware ausgegeben. Wenn der Betrag einer Differenz über diesem kritischen Wert liegt, bezeichnet man die Differenz als signifikant. Man könnte an dieser Stelle noch in die Betrachtung einbeziehen, dass bei der zufallskritischen Absicherung prinzipiell auch die Anzahl der Vergleiche sowie die Richtung des erwarteten Unterschieds beachtet werden sollte. So liegt im Falle eines Kindes mit Lese-Rechtschreib-Schwierigkeiten beispielsweise die Skala AGD der WISC-V mit wesentlich höherer Wahrscheinlichkeit unterhalb als oberhalb von FS. Bei entsprechender Anamnese bzw. Vorabinformationen hat man es also im statistischen Sinne ggf. auch mit einseitigen Hypothesen zu tun, die sich mit entsprechend angepasstem Signifikanzniveau testen lassen. Allerdings möchte ich an dieser Stelle keinen allzu großen Exkurs in die Welt des Hypothesentestens machen.

Vielmehr finde ich bei dieser Art der zufallskritischen Überprüfung von Differenzen einen anderen Aspekt sehr wichtig, der in der Literatur eher selten

thematisiert wird. Dieser Aspekt betrifft eine prinzipielle Annahme, auf der die oben beschriebene Analyse beruht. So geht nämlich die Berechnung der kritischen Differenzen grundsätzlich davon aus, dass der Leistungsbereich der Testperson hierbei keine Rolle spielt. (Dies ist übrigens keine spezifische Schwäche der genannten Testverfahren, sondern vielmehr dem Umstand geschuldet, dass sich bislang noch keine bessere Methode zur leistungsabhängigen Reliabilitätsberechnung durchgesetzt hat.) Leider ist diese Annahme in der Realität jedoch nicht nur manchmal nicht, sondern prinzipiell nie erfüllt. Vielmehr steigt der Anteil der Fehlervarianz an den Messwerten systematisch, je weiter die Leistung einer Testperson vom Durchschnittsbereich entfernt liegt (siehe z. B. Lenhard, Lenhard & Gary, 2019). Dies führt dazu, dass Differenzen in extremen Leistungsbereichen zu häufig als statistisch bedeutsam ausgewiesen werden. Oder um es andersherum zu formulieren: Viele Abweichungen vom intraindividuellen Mittel, die als signifikante Stärken oder Schwächen ausgewiesen werden, kamen in Wirklichkeit doch nur durch Messfehler zustande. Dieser Umstand ist vor allem auch deshalb bedeutsam, weil Intelligenztests in der Praxis überwiegend in Bereichen über- oder unterdurchschnittlicher Leistung eingesetzt werden. Ein Kind mit ganz normaler intellektueller Leistung kommt schließlich eher selten in die Verlegenheit, sich einem Intelligenztest unterziehen zu müssen.

Die Ermittlung von Stärke-Schwäche-Profilen ist also in der Praxis mit großer messtechnischer Unsicherheit behaftet. Das macht sich auch dadurch bemerkbar, dass viele der vermeintlichen Stärken oder Schwächen bei Wiederholungsmessungen mit dem gleichen oder einem ähnlich messenden Verfahren nicht replizierbar sind. Eine klinische Diagnose oder eine gezielte Interventionsmaßnahme auf ein ermitteltes Stärke-Schwäche-Profil zu stützen, stellt demnach ein heikles Unterfangen dar – zumindest sofern das ermittelte Profil nicht durch zusätzliche Evidenz gestützt wird.

Der Validitätsaspekt

Tatsächlich raten viele Wissenschaftler zu großer Zurückhaltung bei der Profilinterpretation auf Indexebene (vgl. z. B. Rost, 2009, S. 160f). Das Hauptargument bezieht sich dabei meistens allerdings gar nicht auf die mangelnde Reliabilität der Profile. Vielmehr wird dabei vor allem auf die Frage abgezielt, ob die einzelnen Indizes überhaupt über den Gesamt-IQ hinaus viel Erklärungs- oder Vorhersagewert für schulische Leistungen oder Nutzen für die Interventionsplanung besitzen. Dies ist eher eine Frage der (prognostischen) Validität als der Reliabilität.

Meines Erachtens liegt in Bezug auf die Verwendung von signifikanten Profilen bei der Diagnose oder Intervention allerdings keine „ob“, sondern eher eine „wann“-Frage vor. Diese Frage lässt sich am einfachsten beantworten, wenn man einmal den Spieß umdreht und fragt, in welchen Fällen der Gesamt-IQ keine gute prognostische Validität besitzt. Stellen Sie sich hierfür ein Auto vor, das von A nach B kommen soll. Die Intelligenz können Sie sich als den Motor vorstellen, der das Auto vorantreibt. Je leistungsfähiger der Motor, desto problemloser und schneller kommt ein Auto in der Regel von A nach B. Wenn Sie mehrere Autos in ein Rennen schicken, dann wird die Leistungsfähigkeit des Motors meistens der ausschlaggebende Faktor dafür sein, welches Fahrzeug am schnellsten ins Ziel kommt. Nun gibt es aber eine Vielzahl an Fällen, in denen eine Fahrt von A nach B trotz leistungsstarkem Motor recht mühsam werden kann. Es könnte beispielsweise sein, dass es keine befestigte Straße gibt, dass also Umweltbedingungen die Fahrt besonders schwierig gestalten. Übertragen auf die Intelligenz bedeutet dies beispielsweise, dass ein Kind, welches in einem bildungsfernen oder sozioökonomisch stark benachteiligten Umfeld aufwächst, sein intellektuelles Potenzial nur schwierig zur Entfaltung bringen kann.

Die Ursachen für mühsames Vorwärtskommen können aber auch im Auto selbst begründet liegen. Stellen Sie sich beispielsweise vor, die Reifen hätten starken Unterdruck oder das Auto hätte eine viel zu große und schwere Last geladen. Es wäre dann unter Umständen zwar noch fahrtüchtig, käme aber nur sehr langsam voran. Auf die Intelligenz übertragen wäre dies z. B. dann der Fall, wenn stark negative emotionale oder motivationale Faktoren die Umsetzung eines intellektuellen Potenzials in eine Leistung verhindern, wie dies bei hoher Leistungsangst oder einer oppositionellen Störung des Sozialverhaltens der Fall sein kann.

Zu guter Letzt gibt es aber auch Schwachstellen des Motors selbst, die die Leistung zum Erliegen bringen können, beispielsweise ein Leck oder Engpass in der Treibstoffzufuhr. Solche „Flaschenhälse“ können dazu führen, dass eine ansonsten leistungsfähige Konstruktion komplett außer Kraft gesetzt wird. Die Schwachstellen können also durch andere Stärken nicht mehr kompensiert werden. Auch in Bezug auf Intelligenz gibt es solche basalen Flaschenhälse. So erschwert z. B. eine stark unterdurchschnittliche Speicherkapazität des auditiven Arbeitsgedächtnisses in erheblichem Maße den Erwerb schulischer Fertigkeiten wie Lesen, Schreiben oder Kopfrechnen. Auch Defizite der Aufmerksamkeitssteuerung können einen ansonsten gut konstruierten intellektuellen Motor nahezu zum Erliegen bringen. In diesem Sinne muss also eine Intelligenzdiagnostik immer auch fragen, ob – und falls ja, wo – solche extrem schwachen Glieder in der Kette existieren (vgl. auch Kubinger, 2009, S. 26).

In Bezug auf intellektuelle Leistungen muss leider konstatiert werden, dass die Schwachstellen, die sich außerhalb des „Motors“ befinden (also emotionale oder motivationale Defizite, sozioökonomische Bedingungen usw.), häufig nicht sonderlich reliabel und valide erfasst werden können. Dem Diagnostiker bzw. der Diagnostikerin bleibt in der Regel nichts anders übrig, als diese Faktoren mit Augenmaß zu bewerten und in das Gesamturteil einfließen zu lassen. Anders sieht es mit den intellektuellen Faktoren aus. Hier kann zuverlässiger quantifiziert werden, ob es sich im statistischen Sinne um extreme Schwachstellen, also

außergewöhnlich starke Senken im Profil, handelt. Dieser grundsätzlichen Frage widmet sich der zweite Schritt, der üblicherweise bei der Analyse von Skalendifferenzen vorgenommen wird. Tatsächlich beruht ja die im Normalfall vorhandene prognostische Überlegenheit des IQ-Wertes gegenüber anderen, enger gefassten Intelligenzfaktoren genau darauf, dass alle intellektuellen Fähigkeiten ein erhebliches Maß an Gemeinsamkeit aufweisen. Umgekehrt bedeutet dies aber, dass der IQ seine prognostische Validität verliert, wenn die intellektuellen Fähigkeiten eines bestimmten Kindes nicht genügend Gemeinsamkeit aufweisen. Im zweiten Analyseschritt wird deshalb getestet, ob die einzelnen Intelligenzfaktoren sich so stark voneinander unterscheiden, dass ein Kind hinsichtlich der Heterogenität seines Intelligenzprofils zu den Extremfällen gehört.

Die praktische Umsetzung

Früher war ja angeblich alles noch viel einfacher auf der Welt – auch die Intelligenztests. Bei der deutschen Urversion des HAWIK (Hardesty & Priester, 1966) gab es beispielsweise anstatt fünf verschiedener Primärindizes nur einen Verbal- und einen Handlungsteil. Hätte man damals untersuchen wollen, ob eine Diskrepanz zwischen diesen beiden Testteilen auffällig hoch ist, so hätte man theoretisch mithilfe der Daten des Handbuchs und der entsprechenden mathematischen Formel (siehe beispielsweise Huber, 1973, S. 139) berechnen können, wie häufig eine so große oder noch größere Diskrepanz überhaupt bei Kindern der Bezugspopulation vorkommt. Für die Berechnung hätte man im Gegensatz zum oben beschriebenen Schritt 1 der Analyse nicht nur die Reliabilitätsindizes der Skalen (damals .92 für den Verbal- und .87 für den Handlungsteil des HAWIK), sondern auch deren Korrelation miteinander benötigt. Diese lag bei etwa $r = .60$. Die Berechnung hätte dann ergeben, dass ein Unterschied von 23 oder mehr IQ-Punkten zwischen Verbal- und Handlungsteil nur bei etwa 5% der Kinder vorkommt. Die 23 Punkte stellten in diesem Sinne also den kritischen Grenzwert für das statistisch auffällige Extremereignis dar.

Ob die 23-Punkte-Regel ihren Ursprung in einer solchen, statistisch möglichst exakten Berechnung beim HAWIK hatte, kann ich aus meiner jetzigen Perspektive nicht mit Sicherheit sagen. Entscheidend ist allerdings, dass die Wahrscheinlichkeit für das Auftreten einer Diskrepanz zwischen zwei Skalen sowohl von der Reliabilität beider Skalen als auch von der Korrelation zwischen den Skalen abhängt. Vor allem letztere ist aber mitnichten bei allen Diskrepanzvergleichen, die in Intelligenztests wie WISC-V oder KABC-II möglich sind, gleich hoch. So korrelieren beispielsweise die Skalen SV und FS in der WISC-V ähnlich hoch miteinander, wie seinerzeit der Verbal- und Handlungsteil beim HAWIK, nämlich zu $r = .59$. Die Korrelation zwischen FS und VG beträgt hingegen nur $r = .26$. Ein Unterschied von 23 Punkten kommt deshalb zwischen FS und VG wesentlich häufiger vor als zwischen FS und SV. Was die Angelegenheit noch ein bisschen komplizierter macht, ist die Tatsache, dass die Wahrscheinlichkeit für einen bestimmten Unterschied zwischen zwei Skalen auch bei dieser Analyse zusätzlich vom Leistungsbereich des Kindes abhängt. So muss beispielsweise für Kinder mit einem IQ von 120 oder mehr Punkten die Skala VG sage und schreibe 41 Punkte unterhalb des G-IQ liegen, damit ein statistisch auffälliges Ereignis mit einer Auftretenswahrscheinlichkeit von unter 5% vorliegt. In der Gruppe der Kinder mit einem IQ unterhalb von 80 Punkten ist das 5%-Kriterium hingegen bereits erreicht, wenn die Skala VG nur läppische 4 Punkte unterhalb des G-IQ liegt.

Das schöne – und das macht die Welt dann wieder ein bisschen einfacher – ist nun, dass ich die eben genannten Wahrscheinlichkeiten nicht mühsam mithilfe komplexer Formeln ausrechnen musste. Vielmehr lassen sie sich anhand der Normstichprobe empirisch ermitteln (was übrigens bei Schritt 1 leider nicht möglich ist). Von den verantwortlichen Psychometrikern wurde dies freundlicherweise gemacht. Die Werte sind in den Testmanualen abgedruckt bzw. werden von der Auswertungssoftware ausgegeben. Bei den Wechsler-Skalen firmieren sie unter der Bezeichnung „Grundraten“, in der KABC-II findet man sie in der Spalte „Selten“ bei der „Analyse der Skalenindices“. Anstatt pauschal die 23-Punkte-Regel anzuwenden, sollte man also lieber – wie bei Schritt 1 – eine Auftretenshäufigkeit festlegen (z. B. 5 %), ab der man eine Diskrepanz als selten – und deshalb extrem – betrachten möchte. In diesem Sinne stellen die Auftretenshäufigkeiten von Diskrepanzen empirisch ermittelte Prüfgrößen für die Nullhypothese dar, dass die Indexskalen bei einem bestimmten Kind genauso viel Gemeinsamkeit aufweisen, wie in der Bezugspopulation, also dass der G-IQ genügend prognostische Validität besitzt. Sinken die Auftretenshäufigkeiten unter die vorher festgelegte Grenze, so stellt der G-IQ vermutlich kein gutes Maß dar, um Schulleistungen, akademische Leistungen oder berufliche Leistungen zu erklären oder zu prognostizieren.

In der Praxis bedeutet dies beispielsweise, dass ein Indexwert von „nur“ 105 Punkten in der Skala VG bei einem Kind mit IQ 130 gar keine außergewöhnliche intraindividuelle Schwachstelle darstellt, da über 15% aller Kinder dieses Leistungsniveaus eine so große oder noch größere Senke in der Skala VG aufweisen. Sollte ein Kind mit einem solchen Profil in der Schule mit Leistungsproblemen ringen, dann würde ich mich als Diagnostikerin lieber auf die Suche nach Schwachstellen machen, die „außerhalb des Motors“ liegen, wie beispielsweise hohe Leistungsangst oder starke motivationale Probleme. Dort müsste dann gegebenenfalls auch eine Intervention ansetzen.

Dass hingegen bei einem G-IQ von 75 in der WISC-V der KLI mehr als 20 Punkte unter dem AFI liegt, kommt nur bei etwa einem Prozent aller Kinder dieses Leistungsbereiches vor. Der extreme Engpass im Arbeitsgedächtnis und/oder in der Verarbeitungsgeschwindigkeit wird hier vermutlich dafür sorgen, dass

der G-IQ die tatsächliche schulische Leistungsfähigkeit eines Kindes deutlich überschätzt. Ein Kind mit einem solchen Leistungsprofil wird sehr viel externe Unterstützung beim schulischen Lernen benötigen. Die Entscheidung über die geeignete Schulform muss also sorgfältig abgewogen werden. Das statistische Zahlenmaterial kann in solchen Fällen dabei helfen, eine vermeintlich „zu niedrige" Schulempfehlung gegenüber Eltern und Behörden zu rechtfertigen.

Verwendete Abkürzungen:

AFI: Allgemeiner Fähigkeitsindex

AGD: Arbeitsgedächtnis

FS: Fluides Schlussfolgern

G-IQ: Gesamt-IQ

HAWIK: Hamburg-Wechsler Intelligenztest für Kinder

KABC: Kaufman Assessment Battery for Children

KLI: Kognitiver Leistungsindex

SV: Sprachverständnis

VG: Verarbeitungsgeschwindigkeit

WISC: Wechsler Intelligence Scale for Children

Literatur:

Amelang, M. & Schmidt-Atzert, Lothar (2006). *Psychologische Diagnostik und Intervention.* Heidelberg: Springer.

Hardesty, F. P. & Priester, H. J. (1966). *Handbuch für den Hamburg-Wechlser-Intelligenztest für Kinder (HAWIK).* Bern: Huber.

Huber, H. P. (1973). *Psychometrische Einzelfalldiagnostik.* Weinheim: Beltz.

Kaufman, A. S., Lichtenberger, E. O., Fletcher-Janzen, E. & Kaufman, N. L. (2005). *Essentials of KABC-II Assessment.* Hoboken (NJ): Wiley & Sons.

Kubinger, K. D. (2009). *Adaptives Intelligenz Diagnostikum 2 (AID 2) - Manual.* Göttingen: Beltz Test GmbH.

Lenhard, A., Lenhard, W. & Gary, S. (2019). Continuous norming of psychometric tests: A simulation study of parametric and semi-parametric approaches. PLoS ONE, 14(9), e0222279. https://doi.org/10.1371/journal.pone.0222279

Rost, D. H. (2009). *Intelligenz – Fakten und Mythen.* Weinheim: Beltz.