BERICHTE

AUS DER

ABTEILUNG ANGEWANDTE PSYCHOLOGIE

Welches sind die nach statistischen Kriterien besten Kurzformen des HAWIK-III?

Urs Schallberger

Nr. 32, 2005

Psychologisches Institut der Universität Zürich
Fachrichtung Angewandte Psychologie
Universitätstr. 84, CH-8006 Zürich

Schallberger, U. (2005). *Welches sind die nach statistischen Kriterien besten Kurzformen des HAWIK-III?* Berichte aus der Abteilung Angewandte Psychologie, Nr. 32. Zürich: Psychologisches Institut der Universität.

Zusammenfassung

Welches sind die nach statistischen Kriterien besten Kurzformen des HAWIK-III?

Es ist eine alte, aber noch nie überprüfte Hypothese, dass für die klassischen Wechsler-Tests eine ganze Reihe von Kurzformen existieren, die eine gute Abschätzung des Gesamt-IQs ermöglichen. Ziel dieses Papiers ist es, sämtliche Kurzformen des HAWIK-III, die eine minimale Anzahl Untertests umfassen und trotzdem rund 90 % der Varianz des Gesamttestwerts erklären, zu identifizieren. Dabei wird auch die Altersabhängigkeit der Güte der IQ-Schätzung berücksichtigt. Als Datenbasis dient die Normierungsstichprobe (N = 1557), welche Kinder im Alter von 6 bis 16 Jahren umfasst. Es zeigt sich, dass in der Gesamtstichprobe und in den oberen Altersgruppen (8 bis 16 Jahre) 4 Untertests genügen, um das Kriterium zu erfüllen. Bei den Jüngsten sind hingegen 5 Untertests notwendig. In vergleichenden Analysen aller möglichen 210 Kurzformen mit 4 und aller möglichen 252 Kurzformen mit 5 Untertests erweisen sich viele Kurzformen als geeignet, was die genannte Hypothese bestätigt. Abschliessend werden die 24 nach statistischen Kriterien besten Vierer-Kurzformen für den gesamten Altersbereich und die 29 besten Fünfer-Kurzformen für die Altersjahre 6 und 7 vorgestellt und einige damit verbundenen Anwendungsprobleme diskutiert.

Abstract

Which short forms of the HAWIK-III are – according to statistical criteria – the best ones?

The goal of the investigation is to identify all short forms of the HAWIK III (the German version of the WISC III) that contain a minimal number of subtests and still explain approximately 90 percent of the variance of the Full Scale IQ score. The precision of the IQ estimate in dependency upon age is also taken into consideration. The normative sample (N = 1557), comprising children aged 6 to 16, serves as the data base. Examination reveals that for the total sample and for older children (aged 8 to 16), 4 subtests are sufficient to fulfill the criterion. For the youngest children (aged 6 to 7), however, 5 subtests are required. The analyses of all possible 210 4-subtest short forms and all possible 252 5-subtest short forms reveal that many short forms are good predictors of the Full Scale IQ score, thus confirming an assumption that was long-suspected but never before tested. The 24 best – according to statistical criteria – 4-subtest short forms for all ages and the 29 best 5-subtest short forms for ages 6-7 are presented and problems associated with their use discussed.

Inhaltsverzeichnis

Zusammenfassung/Abstract 3

1. Einleitung 7
   - 1.1 Ausgangslage und Zielsetzung 7
   - 1.2 Probleme bei der Entwicklung von Kurzformen 9
   - 1.3 Fragestellungen, Untersuchungsstrategie und Vorblick 11
2. Die Untersuchungsstichprobe 12
3. Methoden und Ergebnisse 13
   - 3.1 Schritt 1: Die Bildung von Altersgruppen 13
   - 3.2 Schritt 2: Die Bestimmung der notwendigen Anzahl Untertests 14
   - 3.3 Schritt 3: Vergleichende Analyse aller 210 Vierer-Kurzformen 15
   - 3.4 Schritt 4: Auswahl und Analyse der besten Vierer-Kurzformen 17
   - 3.5 Schritt 5: Kreuzvalidierung der besten Vierer-Kurzformen 20
   - 3.6 Schritt 6: Analyse der Fünfer-Kurzformen für die Altersgruppe 6-7 Jahre 22
4. Bilanz und Hinweise zum Gebrauch der Kurzformen 24

Literatur 27

Anhang 28

1. Einleitung

1.1 Ausgangslage und Zielsetzung

Der Hamburg-Wechsler-Intelligenztests für Kinder (HAWIK) gehört seit seinem ersten Erscheinen (Hardesty & Priester, 1956) zu den meist gebrauchten Individualtests für Kinder. Es handelt sich dabei um eine Adaption der Wechsler Intelligence Scale for Children (WISC; Wechsler, 1949) – ein Instrument, das praktisch weltweit in weitgehend äquivalenten Adaptionen eingesetzt wird und daher den Charakter eines internationalen Standards erhalten hat (vgl. z.B. Georgas, Weiss, van de Vijver & Saklofske, 2003). Bisher sind drei Aktualisierungen vorgenommen worden, die unter den Namen WISC-R (Wechsler, 1974), WISC-III (Wechsler, 1991) und WISC IV (Wechsler, 2004) erschienen sind. Dem WISC-R entspricht der HAWIK-R (Tewes, 1983), dem WISC-III der vor wenigen Jahren erschienene HAWIK-III (Tewes, Rossmann & Schallberger, 1999, 2002).

Eine der vielen Fragen, die in dieser langen Geschichte des HAWIK immer wieder aufgeworfen wurden, ist die Frage nach der Möglichkeit von Kurzformen. Die Durchführung des Tests ist recht zeitaufwändig: Die zehn Untertests, die seit jeher den obligatorischen Kern ausmachen (vgl. die Übersicht in Tabelle 6), erfordern im Normalfall insgesamt rund eine Stunde, bei einzelnen Untersuchungspersonen bzw. klinischen Gruppen auch spürbar mehr. Dazu kommt eine relativ aufwändige Testauswertung. Anderseits ist schon lange bekannt, dass die Untertests des HAWIK (wie bei vielen komplexen Intelligenztests) beträchtliche Interkorrelationen aufweisen. Bei den zehn obligatorischen Untertests des HAWIK-III liegen sie z.B. im Bereich $.21 \leq r \leq .66$; der Median beträgt $r = .43$ (vgl. Tewes et al., 2002, S. 350). *Statistisch gesehen,* weisen die Untertests somit eine beträchtliche Redundanz auf. Dies hat schon früh zu Versuchen geführt, den Gesamt-IQ mit Hilfe einer Regressionsgleichung auf der Basis von deutlich weniger als zehn Untertests abzuschätzen. Die ersten derartigen Kurzformen des HAWIK stammen von Pawlik (1964). Bründler (1984) referierte fünf weitere veröffentlichte Varianten und schlug selber eine Kurzform für den HAWIK-R vor (vgl. auch Bründler, 1989), die unter anderem an einer Kurzform der WISC-R von Kaufman (1976) anknüpfte. Entsprechende Versuche finden sich auch bei anderen Wechsler-Tests (vgl. dazu z.B. Schäuble & Gorlicki, 1998).

Selbstverständlich ist eine Kurzform kein vollwertiger Ersatz für eine vollständige Durchführung des HAWIK. Wechsler hatte bekanntlich die Untertests nicht primär nach statistischen Kriterien ausgewählt. Er hat vielmehr versucht, ein breites Spektrum von Aufgabentypen zusammen zu stellen, die qualitativ möglichst verschiedenartige klinische Beobachtungsmöglichkeiten bieten. Wenn auf die Durchführung eines Teils der Untertests verzichtet wird, entfallen natürlich auch die entsprechenden Beobachtungsmöglichkeiten. Weiter verzichtet man auf die Informationen, die ein vollständiges Testprofil bietet, ebenso auf die Möglichkeit, die Teil-IQ sowie die zusätzliches Indizes zu bestimmen und zu interpretieren (vgl. zu diesen Möglichkeiten Tewes et al., 2002, S. 85 – 94). Und schliesslich handelt es sich

beim gewonnenen IQ lediglich um eine *Schätzung* des normalen HAWIK-IQ, bei der naturgemäss ein grösserer Unsicherheitsbereich in Rechnung zu stellen ist. Es gibt aber durchaus diagnostische Problemstellungen, bei denen diese Einschränkungen in Kauf genommen werden können: Das klarste Beispiel liefern jene Fälle, bei denen es a priori nur um eine relativ grobe Abschätzung des Gesamt-IQ geht, wie es etwa im Rahmen eines Screenings oder eines Forschungsvorhabens der Fall sein kann. Eine Verkürzung der Testdauer kann aber auch zum Beispiel bei wenig belastbaren Untersuchungspersonen angezeigt sein – im Extremfall sogar die einzige Möglichkeit darstellen, überhaupt einen HAWIK durchzuführen. In Tewes et al. (2002, S. 271ff.) finden sich Hinweise auf eine Reihe weiterer diagnostischer Problemstellungen, bei denen der Einsatz einer Kurzform sinnvoll sein kann.

Die regressionsanalytische Natur von Kurzformen bringt es mit sich, dass sie bei jeder Testrevision neu entwickelt werden müssen. Mit dem Erscheinen des HAWIK-III sind damit die bestehenden Kurzformen, die auf dem HAWIK-R basieren, obsolet geworden. *Ziel* der Untersuchungen, die in diesem Papier dargestellt sind, ist es, diese Lücke wieder aufzufüllen und Kurzformen für den HAWIK-III zu entwickeln.[1] Dabei wird aber die Perspektive gegenüber analogen früheren Entwicklungsarbeiten wesentlich ausgeweitet: Ein Charakteristikum dieser früheren Arbeiten ist, dass es jeweils um die Bestimmung einer einzigen "besten" Kurzform und/oder um die Abklärung der Qualität einer oder mehrerer a priori festgelegter Kurzformen ging. Immer wieder wurde dabei die Vermutung geäussert, dass es wahrscheinlich noch eine unbekannte Anzahl anderer Kurzformen gibt, die analoge Eigenschaften zeigen, jedoch beim gewählten Vorgehen nicht identifiziert werden konnten (vgl. schon Pawlik, 1964). Im Folgenden soll diese Vermutung überprüft werden, indem erstmals *alle* Kurzformen, die bestimmte Bedingungen erfüllen, vergleichend analysiert werden.[2] Weiter soll der Frage nachgegangen werden, ob altersgruppenspezifische Kurzformen notwendig sind oder ob allgemein anwendbare Kurzformen vertretbar sind. Bei den meisten früheren Kurzformen konnte diese Frage gar nicht angegangen werden, weil sie auf Grund relativ kleiner und spezieller Stichproben abgeleitet wurden. Die hier berichteten Untersuchungen basieren hingegen auf der gesamten Normierungsstichprobe des HAWIK-III, die für die Beantwortung (auch) dieser Frage sehr gute Voraussetzungen bietet.

Um den Sinn der genannten Zielsetzungen durchsichtiger zu machen und die gewählte Untersuchungsstrategie begründen zu können, sei zunächst ein Überblick über die Probleme gegeben, die sich bei der Entwicklung von Kurzformen für eine Testbatterie wie den HAWIK stellen.

---

[1] Die in diesem Papier vorgestellten Untersuchungen wurden im Jahre 2001 im Hinblick auf die 3., überarbeitete und ergänzte Auflage des Manuals (Tewes et al., 2002) durchgeführt. In Kapitel 7 dieses Manuals findet sich eine Auswahl der hier erarbeiteten Kurzformen, zusammen mit einer sehr summarischen Skizze der Untersuchungen, die ihnen zu Grunde liegen. Das vorliegende Papier liefert die Details nach, ergänzt durch eine Vielzahl weiterer Ergebnisse.

[2] Dabei werden naturgemäss nur jene 10 der total 13 Untertests des HAWIK-III berücksichtigt, die als "obligatorisch" gelten und im Normalfall der IQ-Berechnung zu Grunde zu legen sind.

1.2 Probleme bei der Entwicklung von Kurzformen

Formal gesehen besteht eine Kurzform der hier relevanten Art aus einer Regressionsgleichung, mit deren Hilfe aus den Ergebnissen einer bestimmten Teilmenge von Untertests jener Gesamttestwert abgeschätzt werden kann, der bei der Darbietung aller Untertests resultiert hätte. Geht man von dieser Definition aus, ergeben sich (beim HAWIK-III, mutatis mutandis aber auch bei anderen Testbatterien) im Zusammenhang mit der Entwicklung von Kurzformen folgende Probleme:

Problem 1: Zu prognostizierender Gesamttestwert

Eine grundsätzliche Frage lautet, ob durch die Kurzform der Gesamtrohwert oder der Gesamtnormwert abgeschätzt werden soll. Falls die Normwerte auf einer Lineartransformation der Rohwerte basieren, sind die beiden Vorgehensweisen gleichwertig, und es wäre aus praktischen Gründen (Wegfall der Notwendigkeit, Normentabellen zu benutzen) sinnvoll, direkt den Normwert abzuschätzen. Wenn jedoch die Normierung via Flächentransformation vorgenommen worden ist, könnte – in Abhängigkeit von der Grösse der Abweichung der Rohwert-Normwert-Transformation von der Linearität – mit der direkten Abschätzung des Normwerts ein unerwünschter Präzisionsverlust verbunden sein.

Problem 2: Trade-off zwischen Ökonomie und Präzision

Theoretisch bestehen bei jeder grösseren Testbatterie sehr viele Möglichkeiten für Kurzformen, bei – wie hier – zehn (im Normalfall obligatorischen) Untertests z.B. 1022 Kurzformen, die aus einem bis neun Untertests bestehen. Unter dem Aspekt der Nützlichkeit macht aber eine Kurzform nur dann Sinn, wenn durch eine erhebliche Reduktion der Anzahl Untertests tatsächlich eine spürbare Reduktion der Durchführungsdauer resultiert. Gleichzeitig sollte aber die von der Kurzform gelieferte Schätzung des Gesamttestwerts noch "gut genug" sein. "Gut genug" ist dabei nicht verbindlich definierbar. In der Tradition der Wechsler-Tests wurde manchmal (z.B. Baumert, 1973; Dahl, 1986) eine Mindestkorrelation zwischen Schätzung und Gesamttestwert von mindestens .9 (R^2 = .81) verlangt oder – deutlich strenger –, dass die Schätzung mindestens rund 90 % der Varianz des Gesamttestwerts aufklärt (R^2 = .9; vgl. die Übersichten in Bründler, 1984; Schäuble & Gorlicki, 1998). Daraus resultierten Kurzformen, die aus vier, manchmal auch aus fünf Untertests bestehen.

Problem 3: Bestimmung der Untertests und der Regressionsgleichung

Nach dem eben Gesagten hat die Suche nach einer Kurzform den Charakter eines Optimierungsproblems: Ziel ist es, die kleinste Teilmenge von Untertests zu finden, die den Gesamttestwert mit (noch) als genügend angesehener Präzision abzuschätzen erlaubt. Zur Lösung dieses Problems wird meist die schrittweise Multiple Regression benutzt bzw. empfohlen (z.B. Lienert & Raatz, 1994, S. 356ff.): Dieses Verfahren ermittelt sequentiell, d.h. für k = 1, 2 ... Untertests, jene Kurzform, welche bei gegebenem k die beste Vorhersage des Gesamttestwerts liefert. Der Prozess wird abgebrochen, wenn das Präzisionskriterium (z.B. R^2 = .9)

erstmals erreicht wird.

Die Crux dieses Vorgehens ist seine Stichprobenabhängigkeit: Kleine Zufallsvariationen in der Datenbasis können dazu führen, dass eine bestimmte Teilmenge von Untertests anderen Teilmengen vorgezogen wird (vgl. im Detail z.B. Tabachnick & Fidell, 1996, S. 150ff.). Es kann also neben der auf diese Weise ermittelten Kurzform eine Reihe anderer geben, die eine praktisch äquivalente Vorhersagekraft aufweisen. Die Frage, für welche Untertestkombinationen dies zutrifft, wird durch die schrittweise multiple Regression aber nicht beantwortet. Sie erfordert einen expliziten Vergleich aller in Frage kommenden Kurzformen. Wahrscheinlich wegen der grossen Zahl möglicher Untertestkombination wurde ein solcher expliziter Vergleich noch nie vorgenommen.

Problem 4: Allgemeine Kurzformen vs. Kurzformen für Teilgruppen

Der Anwendungsbereich des HAWIK-III umfasst die elf Altersjahre 6 bis 16 Jahre, wobei die (total 33) Normierungsgruppen auf 4-Monate-Intervallen beruhen. Damit stellt sich die Frage, ob einheitliche Kurzformen für den ganzen Anwendungsbereich oder aber – am andern Extrem – solche für jede Normierungsgruppe oder – als Zwischenlösung – solche für bestimmte Zusammenfassungen von Normierungsgruppen angestrebt werden sollen. Die bisherigen HAWIK-Kurzformen waren pro Altersjahr (Pawlik, 1964), für bestimmte Jahrgangsgruppen (z.B. Bründler, 1989) oder – am häufigsten, aber wohl meist wegen eines für Differenzierungen zu kleinen Stichprobenumfangs – für den ganzen Anwendungsbereich konzipiert.

Offensichtlich handelt es sich auch hier um ein "Trade-off"-Problem: Pragmatisch gesehen wäre es am einfachsten, wenn *eine* Kurzform für den ganzen Anwendungsbereich vorliegen würde. Die Notwendigkeit, von Altersgruppe zu Altersgruppe bestimmte Untertests und/oder die Regressionskoeffizienten auszuwechseln, ist nicht nur eine Fehlerquelle; sie führt auch zu einer eingeschränkten (intra- und interindividuellen) Vergleichbarkeit der Gesamttestwert-Schätzungen. Auf der andern Seite ist es aber auch möglich, dass eine einheitliche Kurzform aus entwicklungspsychologischen Gründen in verschiedenen Altersgruppen Gesamttestwert-Schätzungen von deutlich unterschiedlicher Qualität liefert. Auch dies ist offensichtlich unerwünscht. Das hier zu lösende Problem besteht also darin, einerseits einen möglichst breiten Anwendungsbereich anzustreben, ohne dass anderseits in den verschiedenen Altersgruppen zu grosse Unterschiede in der Qualität der Gesamttestwert-Schätzung in Kauf genommen werden müssen.

Problem 5: Notwendigkeit von Kreuzvalidierungen

Wie weiter oben festgehalten, sind regressionsanalytische Ergebnisse stichprobenabhängig. Sie bedürfen damit der Kreuzvalidierung. Dabei geht es um die Frage, wie gut sich eine in einer Stichprobe gewonnene Regressionsgleichung in einer anderen Stichprobe bewährt. Als Bewährungskriterien stehen dabei verschiedene Kennwerte zur Verfügung (vgl.

z.B. Ang, 1998), die aber im Zusammenhang mit Kurzformen des HAWIK erst ansatzweise benutzt wurden.

1.3 Fragestellungen, Untersuchungsstrategie und Vorblick

Einleitend wurde das Ziel formuliert, alle potentiell in Frage kommenden Kurzformen des HAWIK-III vergleichend zu analysieren, dabei auch deren Altersabhängigkeit zu berücksichtigen und die besten Kurzformen zu identifizieren. Da die eben besprochenen Probleme, die mit dieser Zielsetzung verbunden sind, zum Teil untereinander zusammenhängen, erfordert dieses Ziel ein schrittweises Vorgehen, wobei die Konkretisierung späterer Schritte von den Ergebnissen früherer Schritte abhängig ist. Dies hat zur Folge, dass die Teilfragestellungen dieses Papiers hier lediglich auf einer strategischen Ebene formuliert werden können:

Zunächst sind zwei Grundsatzentscheide zu treffen: Ein erster betrifft den *zu prognostizierenden Gesamttestwert* (vgl. oben, Problem 1). Beim HAWIK-III wurden die IQ-Normen via eine Flächentransformation bestimmt (Tewes et al., 2002, S. 52). Eine (hier nicht wiedergegebene) Analyse der Normentabellen zeigte, dass dabei in den Randbereichen (ca. IQ < 80 und IQ > 120) substantielle Abweichungen von der Linearität entstanden sind. Daher scheint es angebracht, als Schätzwert nicht den IQ zu benutzen, sondern die Wertpunktsumme, die als echte Linearkombination der Untertestergebnisse der Logik der Multiplen Regression besser entspricht. Der zweite Grundsatzentscheid betrifft die Wahl des *Präzisionskriteriums für die Güte der Schätzung* (vgl. oben, Problem 2), die in jedem Fall willkürlichen Charakter hat. Im Folgenden wird im Prinzip vom strengeren der beiden üblichen Kriterien ausgegangen, d.h. von rund 90 % erklärter Varianz ($R^2 \approx .9$, $R \approx .949$). Bei der Auswahl der besten Kurzformen (siehe unten, Schritt 4) wird dieser Grenzwert ganz leicht auf $R^2 \approx .89$ ($R \approx .943$) gelockert. Der Grund für diese Lockerung besteht darin, dass im Hinblick auf die Funktion von Kurzformen eigentlich nicht R^2 die entscheidende Grösse ist, sondern der Standardschätzfehler, und zwar auf der Ebene der IQ-Punkte. Geht man von einem Standardschätzfehler von 5 IQ-Punkten aus, ist auf der Ebene der Wertpunktsumme auch ein $R^2 = .89$ noch tolerierbar.[3]

Ausgehend von diesen Grundsatzentscheiden ergeben sich aus der eingangs genannten Zielsetzung folgende *Teilfragestellungen bzw. Auswertungsschritte:*

Schritt 1: Eine erste empirische Frage ist, ob und welche Altersgruppen für die Analyse der Kurzformen zusammengefasst werden können (vgl. oben, Problem 4).

Schritt 2: Anschliessend ist zu untersuchen, wie viele Untertests notwendig sind, um mindestens rund 90 % der Varianz des Gesamttestwerts aufzuklären (Problem 3).

Schritt 3: Gemäss der leitenden Zielsetzung sind dann alle Kurzformen mit der in

---

[3] Bei der Berechnung des Standardschätzfehlers nach Cohen und Cohen (1983, S. 46ff.) wurde berücksichtigt, dass die Benützung von Normentabellen mit Rundungsfehlern verbunden ist, und dass der IQ wegen der Nichtlinearität der Rohwert-Normwert-Transformation durchschnittlich schlechter prognostizierbar ist als die Wertpunktsumme.

Schritt 2 identifizierten Anzahl Untertests vergleichend zu analysieren. Diese vergleichende Analyse richtet sich primär auf die Güte der Vorhersage des Gesamttestwerts (Problem 3) in der Gesamtstichprobe und in den in Schritt 1 gebildeten Altersgruppen.

Schritt 4: Die obigen Analysen erlauben, die besten Kurzformen zu identifizieren und hinsichtlich verschiedener Aspekte (z.B. Zusammensetzung, Interkorrelationen) zu untersuchen. Das Wort "beste" ist dabei relativ zu verstehen; es bezieht sich hier zudem allein auf statistische Kriterien und auf Grenzwerte, denen eine gewisse Willkürlichkeit zukommt.

Schritt 5: Wie unter Problem 5 besprochen, sind die Regressionsgleichungen der besten Kurzformen auch noch einer Kreuzvalidierung zu unterziehen. Im Hinblick auf die Frage der verantwortbaren Breite des Anwendungsbereichs der Kurzformen (Problem 3) werden als Kreuzvalidierungsgruppen die Altersgruppen nach Schritt 1 benutzt.

Zur Vervollständigung dieses Vorblicks sei noch erwähnt, dass die Ergebnisse der obigen Analyseschritte eine gesonderte Behandlung der Altersjahre 6 und 7 nahe legen werden. Daraus ergibt sich ein *Schritt 6*, bei dem separate Kurzformen für diese Altersjahre geprüft werden. Den Abschluss bilden dann eine zusammenfassende *Bilanz und Hinweise zum Gebrauch der Kurzformen*.

Das skizzierte mehrstufige Vorgehen lässt es als sinnvoll erscheinen, der Übersichtlichkeit halber die benutzten *Methoden* jeweils erst im Zusammenhang mit den *Ergebnissen* der einzelnen Auswertungsschritte zu beschreiben. Der folgende Abschnitt beschränkt sich daher auf eine Darstellung der Untersuchungsstichprobe.

2. Die Untersuchungsstichprobe

Die vorliegende Untersuchung stützt sich auf die Normierungsstichprobe des HAWIK-III (N = 1570), wobei 13 Personen (= 0.8 %) ausgeschlossen wurden, und zwar aus folgenden Gründen: Bei einer Person fehlt das Ergebnis eines der 10 obligatorischen Untertestests. Bei den übrigen 12 Personen handelt es sich um multivariate Outlier, die mit Hilfe des Kriteriums der Mahalanobis-Distanz (Tabachnick & Fidell, 1996, S. 65ff.) identifiziert wurden, und zwar bezogen auf die einzelnen Jahrgangsstichproben. Mit diesen Ausschlüssen soll vermieden werden, dass Personen mit einem sehr ungewöhnlichen Untertestprofil die Ergebnisse der Multiplen Regression über Gebühr beeinflussen.

Eine ausführliche Darstellung der Normierungsstichprobe findet sich in Tewes et al. (2002, S. 47ff.). Sie umfasst Repräsentativstichproben aus Deutschland, der Schweiz und Österreich. Die Altersspanne beträgt – abgedeckt durch eine ungefähre Gleichverteilung – 6 bis 16 Jahre. Beide Geschlechtsgruppen stellen je 50 %. Dies hat sich auch nach der Reduktion um die genannten 13 Personen nicht geändert. Die Verteilung der verbleibenden 1557 Personen auf die Altersjahre ist aus der letzten Zeile von Tabelle 1 ersichtlich. In Bezug auf weitere Stichprobencharakteristika sei auf die genannte Quelle verwiesen.

3. Methoden und Ergebnisse

Die folgende Ergebnisdarstellung orientiert sich an der Abfolge der oben beschriebenen sechs Auswertungsschritte.

3.1 Schritt 1: Die Bildung von Altersgruppen

Die erste empirische Frage richtet sich darauf, ob sich im Hinblick auf die angestrebte vergleichende Analyse von Kurzformen die elf Altersjahre 6 bis 16 Jahre empirisch begründbar in Altersgruppen zusammenfassen lassen. Als Ausgangspunkt für die Beantwortung dieser Frage werden die Korrelationen zwischen den zehn Untertests (Wertpunkte) und der Wertpunktsumme pro Altersjahr benutzt.[4] Bei der gegebenen Ausgangslage (Wertpunktsumme = Summe der Untertestergebnisse) entsprechen diese Korrelationen den Strukturkoeffizienten der Untertests in jener kanonischen Variablen, die in der Wertpunktsumme repräsentiert ist (vgl. Thompson & Borrello, 1985). Sie enthalten damit die für die Bildung von Kurzformen zentrale Information. Die gestellte Frage lässt sich dann auch in der Form ausdrücken, ob es Altersjahre mit ähnlichen Mustern der 10 Untertest-Wertpunktsumme-Korrelationen gibt. Diese Frage wird mit einer hierarchischen Clusteranalyse der nach Fisher z-transformierten Korrelationskoeffizienten angegangen (Ähnlichkeitsmass: quadrierte euklidische Distanz; Clusterungsalgorithmus nach Ward).

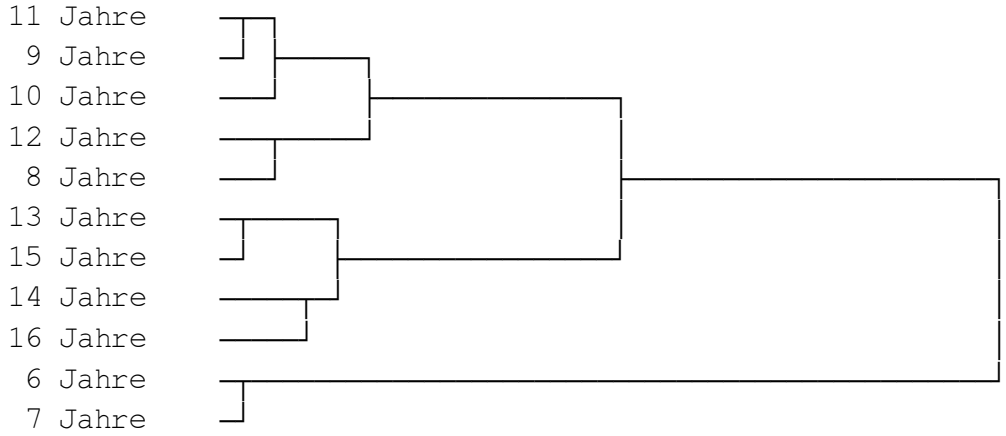

Abbildung 1. Hierarchische Clusteranalyse der nach Fisher z-transformierten Korrelationen zwischen den Untertests und der Wertpunktsumme pro Altersjahr

Das Ergebnis ist in graphischer Form in Abbildung 1 festgehalten. Der Hauptbefund ist, dass sich – wie aus entwicklungspsychologischer Sicht zu erwarten – relativ früh im Clusterungsprozess drei Gruppen benachbarter Altersjahre bilden, nämlich die Gruppen 6-7 Jahre, 8-12 Jahre und 13-16 Jahre. (Die beobachtbaren Abweichungen in den ersten Clusterungsschritten sind vermutlich auf Zufälligkeiten in den Daten zurückzuführen.) Dabei hebt

[4] Diese Korrelationskoeffizienten werden hier aus Raumgründen nicht wiedergegeben. Sie sind aus den relevanten Spalten der Tabellen C1 bis C11 in Tewes et al. (2002) ersichtlich, wobei aber wegen des Ausschlusses der multivariaten Outlier einzelne kleine Abweichungen entstanden sind.

sich die jüngste Gruppe von den zwei älteren viel deutlicher ab als diese zwei sich unterscheiden. Trotzdem scheint es vorsichtig, mit drei Altersgruppen weiterzuarbeiten. Ihre Grösse ist aus der letzten Zeile von Tabelle 1 ersichtlich.

3.2 Schritt 2: Die Bestimmung der notwendigen Anzahl Untertests

Die zweite Frage richtet sich darauf, wieviele Untertests mindestens notwendig sind, um rund 90 % der Varianz der Wertpunktsumme aufzuklären. Wie erinnerlich ist es dann eines der Hauptziele der Untersuchungen, sämtliche mögliche Kurzformen mit dieser Anzahl Untertests einer vergleichenden Analyse zu unterziehen (Schritt 3).

Als Methode zur Beantwortung dieser Frage ist die schrittweise multiple Regression geeignet, die sukzessive, für 1, 2 ... Untertests diejenige Kurzform identifiziert, die jeweils das grösste R^2 liefert. Als Analysegruppen dienen die Gesamtstichprobe und die oben gebildeten Altersgruppen. Zusätzlich werden die einzelnen Altersjahre einbezogen, um zu überprüfen, ob die vorgenommene Zusammenfassung in Altersgruppen wesentliche Unterschiede zwischen den Altersjahren verwischt.

Tabelle 1.
Die Ergebnisse der schrittweisen Regression in der Gesamtstichprobe, in den einzelnen Altersjahren und in den drei Altersgruppen (Anteil erklärter Varianz der Wertpunktsumme)

Anzahl Untertests	Gesamt-stichprobe	pro Altersjahr bzw. Altersgruppe													
		6	7	6-7	8	9	10	11	12	8-12	13	14	15	16	13-16
1	.60	.50	.49	.48	.58	.64	.66	.67	.60	.63	.62	.69	.66	.68	.64
2	.80	.71	.73	.73	.75	.80	.82	.84	.82	.80	.79	.83	.78	.80	.81
3	.86	.82	.81	.82	.84	.88	.91	.89	.88	.87	.87	.88	.88	.86	.87
4	.91	.89	.87	.88	.89	.92	.94	.93	.92	.92	.92	.91	.92	.91	.92
5	.93	.92	.91	.91	.92	.94	.96	.95	.94	.94	.94	.94	.94	.94	.94
6	.95	.94	.94	.94	.95	.96	.97	.97	.95	.96	.96	.95	.95	.96	.96
7	.97	.96	.96	.96	.97	.98	.98	.98	.97	.97	.97	.97	.97	.97	.97
8	.98	.97	.97	.97	.98	.98	.99	.99	.98	.98	.98	.98	.98	.98	.98
9	.99	.99	.99	.99	.99	.99	.99	.99	.99	.99	.99	.99	.99	.99	.99
10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
N	1557	129	140	269	140	141	149	160	147	738	145	134	147	124	550

Anmerkungen: Angegeben sind die adjustierten R^2.

Die Ergebnisse in Tabelle 1 zeigen, dass – wie in der Gesamtstichprobe – in der überwiegenden Mehrzahl der Altersjahre mit vier Untertests die Grenze von 90 % erklärter Varianz erreicht bzw. überschritten wird. Eine Ausnahme bilden die jüngsten Altersjahre (6 bis 8), bei denen dieses Kriterium mit vier Untertests (zum Teil knapp: $.87 \leq R^2 \leq .89$) noch verfehlt und erst mit fünf Untertests erfüllt wird. Die Zusammenfassung in Altersgruppen ändert am gezeichneten Bild nichts Wesentliches. Tabelle 1 liefert auch kein auffälliges Gegenargument gegen die in Abschnitt 3.1 (auf der Grundlage wesentlich differenzierterer Information) vorgenommene Art der Altersgruppenbildung.

3.3 Schritt 3: Vergleichende Analyse aller 210 Vierer-Kurzformen

Die zentrale Frage dieses dritten Auswertungsschrittes ist, ob es neben der einen, in Abschnitt 3.2 identifizierten Vierer-Kurzform mit dem grössten R^2 noch andere gibt, die ein solches vergleichbarer Grössenordnung liefern. Um diese Fragen zu beantworten, werden für sämtliche denkbaren (10 tief 4) 210 Vierer-Kurzformen Multiple Regressionen berechnet, und zwar für die in Abschnitt 3.1 bestimmen Altersgruppen und für die Gesamtstichprobe. Die entsprechenden Verteilungen der R^2 finden sich in Tabelle 2.

Tabelle 2.
Die Verteilung der R^2 (R) aller Vierer-Kurzformen in der Gesamtstichprobe und in den drei Altersgruppen

R^2 (R)	Gesamtstichprobe		6–7 Jahre		8–12 Jahre		13–16 Jahre	
	Häufigkeit	kum. %	Häufigkeit	kum. %	Häufigkeit	kumul. %	Häufigkeit	kum. %
.92 (.959)					1	0.5	1	0.5
.91 (.954)	1	0.5			7	3.8	4	2.4
.90 (.948)	5	2.9			27	16.7	11	7.6
.89 (.943)	21	12.9			47	39.0	27	20.5
.88 (.938)	31	27.6	1	0.5	41	58.6	22	31.0
.87 (.933)	44	48.6	8	4.3	39	77.1	34	47.1
.86 (.927)	39	67.1	9	8.6	13	83.3	30	61.4
.85 (.922)	20	76.7	17	16.7	13	89.5	25	73.3
.84 (.917)	16	84.3	21	26.7	11	94.8	14	80.0
.83 (.911)	7	87.6	32	41.9	4	96.7	13	86.2
.82 (.906)	11	92.9	28	55.2	3	98.1	3	87.6
.81 (.900)	7	96.2	22	65.7	1	98.6	10	92.4
.80 (.894)	2	97.1	28	79.0	2	99.5	3	93.8
.79 (.889)	1	97.6	12	84.8	0	99.5	2	94.8
.78 (.883)	1	98.1	11	90.0	1	100.0	4	96.7
.77 (.877)	1	89.6	5	92.4			2	97.6
.76 (.872)	2	99.5	8	96.2			0	97.6
.75 (.866)	1	100.0	5	98.6			0	97.6
<.75			3	100.0			5	100.0
Total	210		210		210		210	
Maximum R^2 (R)	.906 (.952)		.880 (.938)		.921 (.959)		.919 (.959)	
Oberes Quartil R^2 (R)	.877 (.937)		.837 (.915)		.890 (.944)		.879 (.938)	
Median R^2 (R)	.864 (.930)		.819 (.905)		.879 (.938)		.863 (.929)	
Unteres Quartil R^2 (R)	.846 (.920)		.797 (.893)		.866 (.931)		.843 (.918)	
Minimum R^2 (R)	.747 (.865)		.716 (.846)		.784 (.885)		.698 (.836)	
Schiefe R^2 (s_e)	-1.326 (0.168)		-.497 (0.168)		-1.168 (0.168)		-1.587 (0.168)	
Exzess R^2 (s_e)	2.547 (0.334)		.097 (0.334)		1.889 (0.334)		3.641 (0.334)	

Der auffälligste Befund ist, dass die R^2 generell sehr hoch ausfallen, was auf dem Hintergrund des part-whole-Charakter von Kurzformen und den relativ grossen Interkorrelationen der Untertests nicht ganz unerwartet ist. Werte unter .75 (R = .866) kommen nur vereinzelt vor, die Mediane liegen – ausser bei der bereits auffällig gewordenen Altersgruppe 6-7 Jahre (Median = .819) – im Bereich von rund .86 bis .87. Entsprechend erklären eine ganze

Reihe von Vierer-Kurzformen mindestens 90 % der Varianz der Wertpunktsumme. Am häufigsten ist dies in der Altersgruppe 8-12 Jahre der Fall (35 mal), gefolgt von der Gruppe 13-16 Jahre (16 mal). In der Gesamtstichprobe gilt es – trotz des Einbezugs der Gruppe 6-7 Jahre – immer noch in 6 Fällen. Ginge man vom schwächeren der beiden üblichen Grenzwerte, $R = .9$ ($R^2 = .81$), aus, würden sogar je nach Analysegruppe (inklusive 6-7 Jahre) 66 bis 99 % aller 210 Vierer-Kurzformen das Kriterium erfüllen! Insgesamt ergibt sich damit eine eindrückliche Bestätigung der Vermutung, es gäbe im HAWIK viele Möglichkeiten für gute Kurzformen. Dass es übrigens in allen Gruppen genau einen "Spitzenreiter" gibt, darf nicht überbewertet werden: Der Unterschied zum jeweils nächsten R^2 beträgt nur einige wenige Tausendstel.

Die Ergebnisse in Tabelle 2 implizieren natürlich eine Fülle interessanter Folgefragen, von denen hier nur zwei aufgegriffen werden können:

Eine erste Frage ist, ob es *dieselben* Untertestkombinationen sind, die in den verschiedenen Analysegruppen grössere oder kleinere R^2 liefern. Eine summarische Antwort auf diese Frage findet sich in Tabelle 3, in der für die Menge aller Vierer-Kurzformen die Rangkorrelationen ρ nach Spearman zwischen den R^2 der verschiedenen Untertestkombinationen in den verschiedenen Analysegruppen aufgeführt sind. Die Korrelationen sind alle recht gross ($\rho > .6$). Konsistent mit den Ergebnissen der Clusteranalyse (Abschnitt 3.1) ist die Übereinstimmung zwischen den beiden Altersgruppen 8-12 und 13-16 Jahre ($\rho = .87$) deutlich grösser als jene zwischen diesen beiden Gruppen und der Altersgruppe 6-7 Jahre ($\rho = .73$ bzw. .61). Die bisherige Beobachtung, dass in der Altersgruppe 6-7 Jahre generell niedrigere R^2 resultieren, kann somit dahingehend ergänzt werden, dass in dieser Altersgruppe die Anordnung der Kurzformen nach der Grösse des R^2 etwas anders ausfällt als in den zwei anderen Gruppen. Die Unterschiede der Korrelationen zwischen den drei Altersgruppen einerseits und der Gesamtstichprobe anderseits sind dann eine Folge dieses Unterschieds: Die drei Gruppen stellen ja unterschiedlich grosse Anteile der Gesamtstichprobe und beeinflussen damit das Ergebnis in der Gesamtstichprobe in unterschiedlichem Ausmass.

Tabelle 3.
Die Spearman-Rangkorrelationen der R^2 aller Kombinationen von vier Untertests in den verschiedenen Analysestichproben (n = 210)

	6 – 7 Jahre	8 – 12 Jahre	13 – 16 Jahre
Gesamtstichprobe	.78	.97	.94
6 – 7 Jahre		.73	.61
8 – 12 Jahre			.87

Anmerkung. Alle Koeffizienten sind signifikant von 0 verschieden ($p < .001$)

Eine zweite Folgefrage, die sich bei Tabelle 3 stellt, ist, *welche* Untertests sich *wie* auf die Grösse der R^2 auswirken. Diese Frage würde hier zu weit führen (siehe dazu aber Abschnitt 3.4). Jedoch soll wenigstens abgeklärt werden, ob das Verhältnis von Handlungs- und

Verbaltests eine Rolle spielt, was – wie Tabelle 4 zeigt – in allen Analysegruppen tatsächlich der Fall ist: Die einseitig zusammengesetzten Kurzformen (4 Handlungs- oder 4 Verbaltests) liefern sowohl im Median wie vom erreichten Maximum her gesehen die niedrigsten R^2. Bei den anderen drei Zusammensetzungen (1 Verbal- und 3 Handlungstets; 3 Verbal- und 1 Handlungstest; 2 Verbal- und 2 Handlungstest) sind die Unterschiede geringer. In allen Analysegruppen umfassen aber die Kurzformen mit den grössten R^2 zwei Verbal- und zwei Handlungstests.

Tabelle 4.
Die R^2 der Vierer-Kurzformen nach Anteil Verbal- und Handlungstests

Zusammensetzung:			R^2 Median (Minimum/Maximum)			
Verbal-tests	Handlungs-tests	n	Gesamtstichprobe	6-7 Jahre	8-12 Jahre	13-16 Jahre
0	4	5	.76 (.75/.80)	.75 (.72/.77)	.80 (.78/.84)	.72 (.70/.77)
4	0	5	.82 (.80/.82)	.77 (.74/.78)	.83 (.82/.84)	.81 (.80/.83)
1	3	50	.86 (.78/.89)	.81 (.76/.87)	.87 (.82/.90)	.85 (.73/.90)
3	1	50	.86 (.81/.89)	.82 (.74/.87)	.87 (.84/.90)	.87 (.81/.90)
2	2	100	.87 (.83/.91)	.83 (.76/.88)	.89 (.85/.92)	.87 (.82/.92)
Kruskal-Wallis χ^2 (df = 4)			50.5; $p < .001$	39.5; $p < .001$	49.6; $p < .001$	45.8; $p < .001$

3.4 Schritt 4: Auswahl und Analyse der besten Vierer-Kurzformen

Bilanziert man die bisherigen Ergebnisse im Hinblick auf das einleitend formulierte Ziel, eine Auswahl bester Kurzformen zu treffen, so ergibt sich angesichts der in Tabelle 2 festgehaltenen Ergebnisse eigentlich eine sehr gute Ausgangslage: In der Gesamtstichprobe und in den beiden oberen Altersgruppen je für sich existieren eine ganze Reihe von Kombinationen von vier Untertests, die das in Abschnitt 1.3 eingeführte Kriterium $R^2 \geq .89$ erfüllen. Die Korrelationsanalyse von Tabelle 3 legt zudem nahe, dass es sich dabei in den zwei oberen Altersgruppen weit gehend um dieselben Kurzformen handelt. Hingegen liegen die Verhältnisse in der Altersgruppe 6-7 Jahre etwas anders: Mit vier Untertests kann höchstens 88 % der Varianz des Gesamttestwerts erklärt werden (Tabelle 2), wobei gemäss Tabelle 3 tendentiell etwas andere Untertestkombinationen die grössten Werte erbringen als in den beiden andern Altersgruppen.

Auf diesem Hintergrund scheint es sinnvoll zu sein, für die Altersgruppe 6-7 Jahre spezielle Kurzformen zu bilden, die nicht auf vier, sondern auf fünf Untertests basieren. Dies wird weiter unten (Abschnitt 4.6) geschehen. In Anbetracht der Argumente, die für allgemein verwendbare Kurzformen sprechen (siehe Einleitung, Problem 4), wird diese Gruppe bei der Suche nach den besten Vierer-Kurzformen aber trotzdem mitberücksichtigt, wenn auch mit geringeren Anforderungen an das R^2. Dies führt zur Festsetzung folgender Kriterien: Als beste Vierer-Kurzformen werden jene Kombinationen von vier Untertests bezeichnet,

1.) die in der Gesamtstichprobe ein $R^2 \geq .89$ erreichen

2.) *und* deren altersgruppenspezifische Regressionsgleichungen in den Gruppen 8-12 Jahre und 13-16 Jahre ein $R^2 \geq .89$ *und* in der Altergsruppe 6-7 Jahre ein $R^2 \geq .81$ ergeben

3.) *und* bei denen auch die Anwendung der Regressionsgleichung aus der Gesamtstichprobe in den einzelnen Altersgruppen die obigen Bedingungen erfüllen.

Das Kriterium 2 soll verhindern, dass Kurzformen in die "Bestenliste" Eingang finden, die in einer der drei Gruppen nicht dem gesetzten Minimalstandard genügen, wobei dieser Standard bei der jüngsten Altersgruppe auf $R = .9$ gesenkt wird, der von Autoren früherer Kurzformen schon als untere Grenze für deren Brauchbarkeit benutzt wurde (siehe Einleitung, Problem 2). Kriterium 3 ergibt sich daraus, dass in einer allgemein anwendbaren Kurzform nicht die altersgruppenspezifischen Regressionsgleichungen benutzt werden, sondern jene aus der Gesamtstichprobe, was in den Altersgruppen – *im Vergleich zur altersgruppenspezifischen (d.h. optimalen) Gleichung* – naturgemäss zu einer gewissen Schrumpfung des R^2 führt. Das Kriterium besagt dann, dass auch die geschrumpften R^2 noch das Kriterium 2 erfüllen sollen.

Tabelle 5 enthält die 24 Untertestkombinationen, welche die drei Bedingungen erfüllen (vgl. auch die Anhangstabelle A-1). Grau unterlegt sind jene Kurzformen, die in der jeweiligen Analysegruppe durch die schrittweise Regression als jene mit dem höchsten R^2 identifiziert wurden. Dass sie alle auch in Tabelle 5 aufscheinen, das heisst jeweils auch die andern Auswahlkriterien erfüllen, ist nicht trivial. Weiter zeigt sich, dass die meisten der besten Kurzformen zwei Verbal- und zwei Handlungstests umfassen (vgl. dritte Spalte von Tabelle 5). Immerhin sind aber – v.a. in der zweiten Hälfte der "Rangliste" – auch zwei mit drei und zwei mit nur einem Verbaltest vertreten. Weiter kommen gewisse Untertests sehr viel häufiger vor als andere. Die dritte und vierte Spalte in Tabelle 6 enthalten eine entsprechende Auszählung. Am häufigsten sind der Wortschatztest (88%) und der Mosaiktest (63 %). Eine genaue Inspektion von Tabelle 5 zeigt, dass alle 24 Kurzformen mindestens einen dieser beiden Tests enthalten. Am seltensten treten hingegen der Zahlen-Symbol-Test (8 %) und das Allgemeine Verständnis (17 %) auf.

In Tabelle 5 sind (jeweils nach dem Schrägstrich) auch jene R^2 aufgeführt, die sich ergeben, wenn statt der altersspezifischen Regressionsgleichung jene aus der Gesamtstichprobe angewendet wird. Es zeigt sich, dass diese Substitution in aller Regel auf die zweistelligen R^2 kaum eine Auswirkung hat (Range der Schrumpfung: -.0002 bis -.0074). Es ist denn auch so, dass das Kriterium 3 zum Ausschluss keiner einzigen Kurzform führte (vgl. auch die differenziertere Schrumpfungsanalyse in Abschnitt 3.5).

Tabelle 5.
Die 24 besten Vierer-Kurzformen (geordnet nach der erklärten Varianz in der Gesamtstichprobe).

Nr.	Untertests	Anzahl Verbaltests	R^2 (Gleichung aus Altersgruppe/aus Gesamtstichprobe)			
			Gesamt	6-7 Jahre	8-12 Jahre	13-16 Jahre
4-1	AW AV BO MT	2	.91	.85/.85	.92/.92	.91/.91
4-2	AW WT BO MT	2	.90	.85/.85	.91/.91	.92/.92
4-3	GF WT BO MT	2	.90	.83/.83	.91/.91	.91/.91
4-4	RD WT BO MT	2	.90	.84/.84	.91/.91	.90/.90
4-5	AW WT MT FL	2	.90	.86/.86	.90/.90	.90/.90
4-6	RD WT BE FL	2	.90	.88/.88	.91/.91	.89/.89
4-7	GF RD WT FL	3	.89	.86/.86	.90/.90	.90/.90
4-8	AW WT ZS MT	2	.89	.84/.84	.90/.90	.90/.90
4-9	GF WT BO FL	2	.89	.87/.86	.90/.90	.90/.90
4-10	AW WT BE MT	2	.89	.84/.84	.90/.90	.91/.91
4-11	RD WT BO FL	2	.89	.87/.86	.90/.90	.89/.89
4-12	GF WT MT FL	2	.89	.85/.85	.90/.90	.90/.90
4-13	RD WT MT FL	2	.89	.86/.86	.90/.90	.90/.89
4-14	RD WT BE MT	2	.89	.86/.85	.90/.90	.90/.89
4-15	GF AV BO MT	2	.89	.84/.83	.90/.90	.90/.89
4-16	WT AV BO MT	2	.89	.84/.84	.91/.91	.89/.89
4-17	AW WT BO FL	2	.89	.87/.87	.90/.90	.89/.89
4-18	WT BE BO MT	1	.89	.83/.83	.90/.90	.90/.89
4-19	GF WT BE FL	2	.89	.86/.86	.90/.90	.89/.89
4-20	WT BO MT FL	1	.89	.85/.84	.90/.90	.89/.89
4-21	AW WT BE FL	2	.89	.87/.87	.89/.89	.89/.89
4-22	AW RD WT FL	3	.89	.86/.86	.90/.89	.89/.89
4-23	AW AV BE MT	2	.89	.85/.84	.89/.89	.89/.89
4-24	GF WT ZS FL	2	.89	.83/.83	.90/.90	.89/.89

Anmerkungen: Grau unterlegt ist jeweils der Maximalwert der entsprechenden Analysestichprobe. Die gestrichelte Linie trennt jene sechs Kurzformen vom Rest ab, die in der Gesamtstichprobe ein $R^2 \geq .9$ erreichen.

Tabelle 6.
Die Häufigkeit des Vorkommens der Untertests in den 24 besten Vierer-Kurzformen (Tabelle 5) und den 29 besten Fünfer-Kurzformen (Tabelle 9)

Subtest mit Abkürzung		Häufigkeit in Vierer-Kurzformen		Häufigkeit in Fünfer-Kurzformen	
		absolut	in %	absolut	in %
Allgemeines Wissen	AW	9	38 %	12	41 %
Gemeinsamkeitenfinden	GF	7	29 %	11	38 %
Rechnerisches Denken	RD	7	29 %	21	72 %
Wortschatztest	WT	21	88 %	20	69 %
Allgemeines Verständnis	AV	4	17 %	14	48 %
Bilderergänzen	BE	7	29 %	14	48 %
Zahlen-Symbol-Test	ZS	2	8 %	7	24 %
Bilderordnen	BO	11	46 %	10	35 %
Mosaik-Test	MT	15	63 %	18	62 %
Figurenlegen	FL	13	54 %	18	62 %
Erwartungswert			40 %		50 %

Die 24 Kurzformen der Tabelle 5 können als eine Art Parallelformen betrachtet werden. Zwar enthalten sie teilweise dieselben Subtests; jedoch werden diese in den verschiedenen Untertestkombinationen jeweils unterschiedlich gewichtet. Auf diesem Hintergrund soll noch kurz auf ihre total (23 x 24)/2 = 276 *Interkorrelationen* eingegangen werden. Die entsprechenden Koeffizienten bewegen sich in einem Range zwischen .82 und .97 (Median = .92). Die überwiegende Mehrzahl, nämlich 75 %, erreichen mindestens .90. Als Hauptdeterminante der unterschiedlichen Grösse erweist sich die Anzahl gemeinsamer Untertests, die von 0 bis 3 variiert. Die Rangkorrelation nach Spearman zwischen dieser Anzahl und der Korrelation beträgt .95 ($p < .001$). Bestimmt man weiter die Mediane der Korrelationen für die 24 Kurzformen, ergeben sich für 22 Formen Werte von .92 und .93. Lediglich die beiden Kurzformen 4-15 und 4-23 fallen mit einem Median von .88 ganz leicht ab. Sie haben am wenigsten Untertests mit der Gesamtheit der jeweils übrigen Kurzformen gemeinsam. Legt man ein nicht zu strenges Kriterium an, spricht also nichts dagegen, die 24 Kurzformen als (Quasi-) Parallelformen zu betrachten.

3.5 Schritt 5: Kreuzvalidierungen der besten Vierer-Kurzformen

In diesem Schritt ist nun noch zu überprüfen, ob die 24 besten Kurzformen auch einer Kreuzvalidierung standhalten. Methodisch wird dabei nach der Logik der sog. doppelten Kreuzvalidierung vorgegangen, d.h. es wird untersucht, in welchem Ausmass die in einer Stichprobe A bestimmte Gleichung in einer Stichprobe B ähnliche Schätzungen liefert wie die Gleichung, die aus der Stichprobe B selbst stammt, also optimal ist, und umgekehrt. Dabei werden zwei Kriterien benützt: die beim genannten Vergleich resultierende Schrumpfung des R^2 und der sog. Invarianzkoeffizient (Ang, 1998), der im Hinblick auf die diagnostische Funktion von Kurzformen ebenfalls wichtig erscheint. Dieser Koeffizient (eine Korrelation) beschreibt, in welchem Ausmass die Anordnung der Untersuchungspersonen durch die beiden Schätzgleichungen übereinstimmt. Als Stichproben A bzw. B dienen die drei Altersgruppen, ferner – nicht als echte Kreuzvalidierung, sondern zur Kontrolle der Übertragbarkeit der Gleichung aus der Gesamtstichprobe auf die Altersgruppen und umgekehrt – auch die Gesamtstichprobe. Bezieht man die geschilderte Logik der doppelten Kreuzvalidierung auf alle möglichen Paare dieser vier Stichproben, ergeben sich pro Kurzform je 12 Kennwerte der Schrumpfung und der Invarianz. Aus Raumgründen sind in Tabelle 7 lediglich die Kennwerte der vier ersten und vier letzten Kurzformen aus Tabelle 5 zusammengestellt. Die Ergebnisse der Beispiele sind für alle Kurzformen typisch: Das Ausmass der Schrumpfung ist in der Regel ausserordentlich klein und liegt in den meisten Fällen deutlich unter 0.005. Die (sehr seltenen) Werte über .005 treten zudem vorwiegend dann auf, wenn die Altersgruppe 6-7 Jahre die "Herkunfts-" oder "Anwendungsstichprobe" einer Gleichung darstellt. Ein analog günstiges Bild liefern die Invarianzkoeffizienten, die sich meist zwischen .999 und 1.0 bewegen; niedrigere Koeffizienten (extrem selten unter .998) finden sich wiederum praktisch nur im Zusammenhang mit der Altersgruppe 6-7.

Tabelle 7.
Kreuzvalidierungen ausgewählter Vierer-Kurzformen

4er-Kurzform in der Stichprobe:	Schrumpfung (R^2)				Invarianz (r)			
	6-7	8-12	13-16	alle	6-7	8-12	13-16	alle
4-1: 6-7 Jahre	.852	.004	.006	.004	-	.997	.996	.998
8-12 J.	.004	.920	.000	.000	.998	-	1.0	1.0
13-16 J.	.005	.000	.908	.000	.997	1.0	-	1.0
alle	.003	.000	.000	.906	.998	1.0	1.0	-
4-2: 6-7 Jahre	.846	.001	.001	.001	-	.998	.999	.999
8-12 J.	.002	.912	.002	.000	.999	-	.999	1.0
13-16 J.	.002	.002	.920	.000	.999	.999	-	1.0
alle	.001	.000	.000	.904	.999	1.0	1.0	-
4-3: 6-7 Jahre	.830	.004	.002	.000	-	.998	1.0	1.0
8-12 J.	.002	.908	.002	.002	.999	-	.999	1.0
13-16 J.	.002	.004	.912	.002	1.0	.999	-	1.0
alle	.000	.000	.000	.898	1.0	1.0	1.0	-
4-4: 6-7 Jahre	.841	.004	.004	.002	-	.998	.998	.999
8-12 J.	.004	.910	.004	.002	.999	-	.998	1.0
13-16 J.	.003	.005	.904	.001	.998	.997	-	.999
alle	.002	.002	.002	.897	.999	.999	.999	-
4-21: 6-7 Jahre	.896	.003	.006	.003	-	.998	.996	.999
8-12 J.	.002	.894	.001	.000	.999	-	.999	1.0
13-16 J.	.005	.001	.886	.001	.997	.999	-	1.0
alle	.002	,000	.001	.885	.999	1.0	1.0	-
4-22: 6-7 Jahre	.858	.000	.005	.001	-	1.0	.997	.999
8-12 J.	.000	.895	.002	.000	1.0	-	.999	1.0
13-16 J.	.004	.003	.888	.001	.998	.999	-	.999
alle	.001	.000	.001	.885	.999	1.0	.999	1.0
4-23: 6-7 Jahre	.846	.002	.004	.001	-	.999	.998	.999
8-12 J.	.001	.895	.003	.001	.999	-	,998	1.0
13-16 J.	.004	.004	.888	.001	.998	.998	-	1.0
alle	.001	.001	.001	.885	.999	1.0	.999	-
4-24: 6-7 Jahre	.831	.004	.006	.003	-	.998	.996	.998
8-12 J.	.003	.895	.001	.000	.998	-	.999	1.0
13-16 J.	.006	.001	.894	.001	.997	.999	-	1.0
alle	.003	.000	.001	.885	.998	1.0	1.0	-

Anmerkungen: Erste Spalte: Nummer der Kurzform (vgl. Tabelle 5) und Berechnungsstichprobe. Spalte "Schrumpfung *(R^2)*": Betrag der Differenz zwischen dem R^2 aus der Gleichung aus der in der Spaltenüberschrift angegebenen Stichprobe und dem Referenzwert (unterstrichen). Spalte "Invarianz *(r)*": Korrelation zwischen der Schätzung aus der Gleichung der in der Spaltenüberschrift angegebenen Stichprobe und der Schätzung aus der Gleichung aus der Analysestichprobe.

Sowohl die Schrumpfungs- wie die Invarianzanalyse liefern somit Ergebnisse, die für eine weitgehende Austauschbarkeit der Schätzungen sprechen, die aufgrund der Regressions-

gleichungen aus den verschiedenen Stichproben vorgenommen wurden. Dies gilt insbesondere auch für die Gleichung aus der Gesamtstichprobe, die aber den Vorteil hat, auf dem grössten Stichprobenumfang zu basieren.

3.6 Schritt 6: Analyse der Fünfer-Kurzformen für die Altersgruppe 6-7 Jahre

Wie einleitend zu Schritt 4 in der Zwischenbilanz festgehalten wurde, drängen sich für die Altersgruppe 6-7 Jahre spezielle Kurzformen auf, die fünf Untertests umfassen. Daher wurden – nach derselben Strategie wie die Vierer-Kurzformen – bei dieser Altersgruppe auch sämtliche 252 möglichen Fünfer-Kurzformen analysiert. Aus Raumgründen seien hier nur einige Hauptergebnisse festgehalten:

Tabelle 8.
Die Verteilung der R^2 aller Fünfer-Kurzformen in der Altersgruppe 6-7 Jahre

R^2 (R)	Häufigkeit	kumul. %
.91 (.954)	5	2.0
.90 (.948)	24	11.5
.89 (.943)	39	27.0
.88 (.938)	49	46.4
.87 (.933)	53	67.5
.86 (.927)	31	79.8
.85 (.922)	20	87.7
.84 (.917)	18	94.8
.83 (.911)	5	96.8
.82 (.906)	3	98.0
.81 (.900)	1	98.4
.80 (.894)	4	100.0
Total	252	
Maximum R^2 (R)	.914 (.956)	
Oberes Quartil R^2 (R)	.886 (.941)	
Median R^2 (R)	.873 (.934)	
Unteres Quartil R^2 (R)	.858 (.926)	
Minimum R^2 (R)	.796 (.892)	
Schiefe R^2 (s_e)	-.807 (0.153)	
Exzess R^2 (s_e)	1.055 (1.218)	

Die Verteilung der R^2 der 252 möglichen Fünfer-Kurzformen in der Altersgruppe 6-7 Jahre (vgl. Tabelle 8) zeigt nun ein wesentlich besseres Bild als das Analogon in Tabelle 2: Alle 252 Fünfer-Kurzformen erreichen ein R^2 von mindestens .8. Das oben angewandte Kriterium von .89 wird von 58 (27 %) erfüllt! Jene 29 Untertestkombination, die ein R^2 von mindestens .9 erreichen, sind in Tabelle 9 aufgeführt. Dieser Tabelle ist zu entnehmen, dass alle diese Kurzformen entweder zwei oder drei Verbaltests bzw. Handlungstests umfassen, die extremen Verhältnisse 4 : 1 oder 5 : 0 also nicht vorkommen. Wie bei den Vierer-Kurzformen

kommen die verschiedenen Untertests unterschiedlich häufig vor (vgl. fünfte und sechste Spalte in Tabelle 6). Der grösste Unterschied besteht darin, dass in den Fünfer-Kurzformen für die Altersgruppe 6-7 Jahre das Rechnerische Denken eine sehr viel bedeutsamere Rolle spielt als in den Vierer-Kurzformen für die Gesamtstichprobe und bei den jüngsten Kindern eine dem Wortschatztest analoge Bedeutung hat.

Tabelle 9.
Die 29 Kurzformen, die in der Altersgruppe 6-7 Jahre mindestens 90 % der Varianz erklären (absteigend angeordnet nach R^2)

Nr.	Untertests	Anzahl Verbaltests	R^2
5-1	RD WT BE BO FL	2	.91
5-2	AW WT BE BO FL	2	.91
5-3	GF RD WT BE FL	3	.91
5-4	GF WT BE BO FL	2	.91
5-5	AW RD WT BE FL	3	.91
5-6	RD WT BE MT FL	2	.90
5-7	GF RD AV ZS MT	3	.90
5-8	AW RD AV ZS MT	3	.90
5-9	GF RD WT BO FL	3	.90
5-10	AW GF WT BO FL	3	.90
5-11	AW WT BO MT FL	2	.90
5-12	RD WT AV MT FL	3	.90
5-13	GF RD AV BE MT	3	.90
5-14	RD AV BE ZS MT	2	.90
5-15	RD WT AV BE MT	3	.90
5-16	AW RD WT MT FL	3	.90
5-17	RD WT BE ZS FL	2	.90
5-18	AW RD WT BO FL	3	.90
5-19	AW WT BE MT FL	2	.90
5-20	AW RD AV BE MT	3	.90
5-21	RD WT AV BE FL	3	.90
5-22	GF RD WT MT FL	3	.90
5-23	AW AV ZS BO MT	2	.90
5-24	AW GF WT BE FL	3	.90
5-25	RD WT AV BO MT	3	.90
5-26	RD WT AV ZS MT	3	.90
5-27	AW GF AV ZS MT	3	.90
5-28	GF RD AV BO MT	3	.90
5-29	GF RD AV MT FL	3	.90

Die 406 Interkorrelationen der 29 Kurzformen liegen ungefähr in derselben Grössenordnung wie jene der Vierer-Kurzformen (Median = .93, Minimum = .81, Maximum = .98) und zeigen eine ähnliche Verteilung. Ebenfalls die Kreuzvalidierung, die auf Grund einer Zufallshalbierung der Stichprobe der Gruppe 6-7 Jahre vorgenommen wurde, liefert Ergebnisse,

die mit jenen in Tabelle 7 vergleichbar sind. Dies war aufgrund der *relativ* geringen Grösse der beiden Stichprobenhälften ($N \approx 135$) nicht ohne weiteres zu erwarten. Offenbar wirkt sich jedoch die Tatsache, dass es sich um *Zufalls*hälften (und nicht um verschiedene Altersgruppen wie in Tabelle 7) handelt, kompensierend aus.

4. Bilanz und Hinweise zum Gebrauch der Kurzformen

Hauptziel der dargestellten Untersuchungen war, für den HAWIK-III Kurzformen zu entwickeln. Auf dem Hintergrund der alten (z.B. Pawlik, 1964), aber noch nie überprüften Vermutung, dass es eine ganze Reihe "guter" Kurzformen gebe, galt das Interesse nicht nur der Identifikation *einer* besten Kurzform oder der Überprüfung einzelner, a priori definierter Untertestkombinationen. Vielmehr sollten alle möglichen Kurzformen mit einer bestimmten Anzahl Untertests vergleichend analysiert werden, um auf diese Weise sämtliche "beste" Kurzformen zu identifizieren. Weiter wurde unter verschiedenen Gesichtspunkten der Frage nachgegangen, ob altersspezifische Kurzformen notwendig sind oder ob – was praktische Vorteile aufweist – allgemein anwendbare Kurzformen vertretbar sind. Dies alles geschah auf dem Hintergrund einer Auflistung der Probleme bei der Entwicklung von Kurzformen und in mehreren Untersuchungsschritten. Für die detaillierten Überlegungen und Befunde muss auf die früheren Abschnitte verwiesen werden. Hier seien nochmals die Hauptergebnisse zusammengefasst und einige Fragen diskutiert, die den praktischen Gebrauch der hier vorgestellten Kurzformen betreffen.

Ein zentraler Punkt bei der Frage nach besten oder guten Kurzformen ist natürlich die Festlegung eines Qualitätskriteriums. Wie in der Einleitung erwähnt, wird in der Tradition der Wechsler-Tests oft davon ausgegangen, dass eine Kurzform bei der regressionsanalytischen Abschätzung des Gesamttestwerts ein R^2 von rund .9 liefern soll. Im Rahmen dieses Papiers wird dieses Kriterium zum Teil durch das im Hinblick auf die diagnostische Zielsetzung einer Kurzform näher liegende des Standardschätzfehlers ersetzt. Dabei wird von einem Kriterium von 5 IQ-Punkten ausgegangen, was einem $R^2 \geq .89$ entspricht.

Ein erstes Hauptergebnis ist nun, dass sich unter dem Aspekt der Kurzformproblematik die Verhältnisse in der Altersgruppe 6 bis 7 Jahre und jenen im Rest des Anwendungsbereichs des HAWIK (8 bis 16 Jahre) in verschiedenen Hinsichten unterscheiden. Der im gegebenen Zusammenhang wichtigste Unterschied ist, dass im Altersbereich 8 bis 16 Jahre (und auch in der Gesamtstichprobe) vier Untertests genügen, um die genannten Qualitätskriterien zu erfüllen, während dies in der Altersgruppe 6 bis 7 Jahre erst mit fünf Untertests möglich ist.

Bei dieser Ausgangslage wurde bei der Suche nach den besten Kurzformen eine doppelte Strategie gefahren: Zum einen wurden alle 210 möglichen Vierer-Kurzformen auf ihre Tauglichkeit für den gesamten Anwendungsbereich des HAWIK geprüft, wobei aber bei der

Altersgruppe 6-7 Jahre notgedrungen das Qualitätskriterium zu lockern war ($R \geq .9$). Zum andern wurden speziell für den Anwendungsbereich 6-7 Jahre sämtliche 252 Fünfer-Kurzformen vergleichend untersucht. In beiden Fällen ergaben sich eine ganze Menge brauchbarer Kurzformen, deren Schätzgleichungen sich in den Kreuzvalidierungen als ausserordentlich robust erwiesen.[5] Die 24 besten Vierer-Kurzformen für die Gesamtstichprobe, deren Schätzgleichungen und Standardschätzfehler (total und pro Altersgruppe) sind in der Anhangstabelle A-1 enthalten. Bei den Fünfer-Kurzformen für die Gruppe 6-7 Jahre ergaben sich nach den analogen Kriterien sogar 58 beste Kurzformen. Die Standardschätzfehler und Schätzgleichungen jener 29, die mindestens 90% der Varianz der Wertpunktsumme erklären, sind aus der Anhangstabelle A-2 ersichtlich.

Insgesamt liefern diese Befunde somit eine eindrückliche Bestätigung der alten Vermutung, es gebe im HAWIK viele Untertestkombinationen, die eine gute Abschätzung des Gesamttestwerts ermöglichen. Dass diese Vermutung bisher noch nie überprüft wurde, dürfte mit dem dafür notwendigen enormen Rechenaufwand zusammenhängen. Während für die Identifikation einer besten Kurzform lediglich *eine* (schrittweise) multiple Regression notwendig ist, stecken allein hinter den vergleichenden Analysen der Tabellen 2 und 8 bereits rund 1'100 (Standard-) Multiple Regressionen.

Aus der Vielzahl der gefundenen, gemäss der untersuchten Kriterien guten Kurzformen entsteht nun aber für die Praxis ein neues Problem, nämlich die Frage, welche der vielen Kurzformen denn zu benutzen sei. Für eine fundierte Antwort auf diese Frage wäre es notwendig, die verschiedenen Kurzformen hinsichtlich ihrer formalen und diagnostischen Eigenschaften noch weitergehend zu studieren als es hier bei dieser ersten Übersicht möglich war. Natürlich steht es aber beispielsweise frei, analog zu den Kurzformen früherer HAWIK-Versionen, die auf einer schrittweisen multiplen Regression basierten, die Kurzformen auf dem Rangplatz 1 der Tabellen A–1 und A–2 zu wählen. Ebenso ist es möglich, auf Grund der zusätzlichen Informationen in diesen Tabellen eigene statistische Kriterien für eine beste Kurzform zu formulieren. Zusätzlich dürfte es aber in jedem Falle sinnvoll sein, auch nicht-statistische Kriterien (mit) zu berücksichtigen. So hat zum Beispiel Kaufmann (1976), ein hervorragender Kenner der Wechsler-Tests, für den WISC-R unter anderem aus inhaltlichen Überlegungen die Version RD WT BO MT (4-4 in Tabelle 5) empfohlen. Bei einer Umfrage bei einer kleinen Gruppe von Schweizer Diagnostikern mit langjähriger HAWIK-Erfahrung resultierte eine Präferenz für die Version GF WT BO MT (4-3 in Tabelle 5). Dass beide Versionen auch nach den hier berichteten statistischen Analysen zu den bestmöglichen Kurzversionen für die Gesamtstichprobe gehören, ist bemerkenswert. Und schliesslich ist auch gut denkbar, dass spezifische klinische Erfahrungen mit den verschiedenen Untertests in einem bestimmten Arbeitsbereich in die Überlegungen einbezogen werden könnten. Wie die Unter-

---

[5] Dieser bei der Multiplen Regression – wie früher erwähnt – seltene Sachverhalt erklärt sich aller Wahrscheinlichkeit nach durch die grossen Stichprobenumfänge, die hier zur Verfügung standen.

testauswahl auch immer ausfällt: Mit praktisch bedeutsamen Implikationen für die Präzision der IQ-Schätzung ist nicht zu rechnen, solange es um Problemstellungen geht, bei denen der Einsatz einer Kurzform sinnvoll ist. Alle in den Tabellen 5 und 9 aufgeführten Kurzformen sind ja mit $r \geq .94$ sehr gute Prädiktoren der auf Grund sämtlicher Untertests bestimmten Wertpunktsumme.

Im Altersbereich 6-7 Jahre stellt sich das zusätzliche Problem, ob eine Vierer- oder eine Fünfer-Kurzform verwendet werden soll. Dies ist vor allem eine Frage der angestrebten Güte der IQ-Schätzung. Wie der Vergleich der Tabellen A-1 und A-2 zeigt, sind die Standardschätzfehler der Fünfer-Kurzformen deutlich kleiner als jene der Vierer-Kurzformen. Im Hinblick auf einen Einsatz einer Fünfer-Kurzform, bei dem (z.B. für Quervergleiche zwischen Altersgruppen) auch das Ergebnis einer Vierer-Kurzform interessiert, sei noch darauf aufmerksam gemacht, dass verschiedene Fünfer-Kurzformen Erweiterungen von Vierer-Kurzformen um einen zusätzlichen Untertest darstellen (vgl. Hinweise in Tabelle A-2).

Abschliessend sei nochmals daran erinnert, dass in der diagnostischen Praxis ein Kurzform-IQ deutlich vorsichtiger interpretiert werden sollte, als es bei Testergebnissen ohnehin angezeigt ist. Dafür sprechen vor allem folgende Gründe: Wie schon früher betont, handelt es sich bei einem Kurzform-IQ um eine *Schätzung* des normalen HAWIK-IQ. Gegenüber dem normalen HAWIK-IQ beruht er auf wesentlich weniger Untertests, ist also empfindlicher für diagnostisch irrelevante Störungen der Leistung eines Kindes und/oder Durchführungs- und Auswertungsfehlern des Testleiters in einem einzelnen Untertest. Weiter sind die Standardschätzfehler in der bekanntlich ohnehin überdifferenzierten IQ-Norm doch – wie schon bei früheren Kurzformen – recht beträchtlich und damit die Unsicherheitsbereiche recht gross. Geht man grob von einem Standardschätzfehler von 5 IQ-Punkten aus, beträgt das Konfidenzintervall bei 68%iger Sicherheit rund ± 5 Punkte und jenes bei 95%iger (bzw. 99%iger) Sicherheit rund ± 10 (bzw. 13) Punkte.[6] Und schliesslich ist noch daran zu erinnern, dass die entscheidende Grundlage von Kurzformen, die Regressionskoeffizienten, nicht etwa "Naturkonstanten" sind. Sie sind vielmehr Ausdruck des Interkorrelationsmusters der Untertests in der Eichstichprobe. Dieses Muster kann aber in speziellen Gruppen z.B. regionaler oder psychopathologischer Art mehr oder weniger variieren. Die Anwendung einer der Gleichungen aus den Anhangstabellen kann daher in einer solchen Gruppe zu einer systematischen Unter- oder Überschätzung des IQ führen. Nur wenn man den auch in dieser Tatsache zum Ausdruck kommenden Charakter eines Kurzform-IQ als *Schätzung* mitberücksichtigt, kann die Anwendung einer Kurzform bei den einleitend genannten diagnostischen Problemstellungen ein nützliches Hilfsmittel darstellen.

---

[6] Die exakten Konfidenzintervalle sind an sich auch von der Abweichung der individuellen Wertpunkte vom Mittelwert 10 abhängig (vgl. z.B. Cohen & Cohen, 1983, S. 111ff.). Die Grösse der entsprechenden Korrekturen hängt aber zentral von der Grösse jener Stichprobe ab, in der die Regressionskoeffizienten gewonnen wurden. Bei den Stichprobenumfängen, auf denen die hier präsentierten Kurzformen basieren, handelt es sich dabei um Werte, die selbst bei extremen Wertpunkten die Grösse von 0.1 IQ-Punkten nicht überschreiten.

Literatur

Ang, R. P. (1998). Use of double cross-validation and bootstrap methods to estimate replicability of results of multiple regression. Perceptual and Motor Skills, 86, 1143-1152.

Baumert, I. (1973). Untersuchungen zur diagnostischen Valenz des HAWIK und die Entwicklung einer Kurzform. Bern: Huber.

Bründler, M. (1984). Eine Kurzform für den HAWIK-R. Berichte aus der Abteilung Angewandte Psychologie, Nr. 21. Zürich: Psychologisches Institut der Universität.

Bründler, M. (1989). HAWIK-R-Kurzform für 8-9jährige und 10-12jährige Kinder. Unterlage zu den Informationstagen der Schweizer Kinder- und Jugendpsychologen. Solothurn: 7.-9.März 1989.

Cohen, J. & Cohen, P. (1983). Applied multiple regression/correlation analysis for the behavioral sciences. Hillsdale: Erlbaum.

Dahl, G. (1986). WIP - Reduzierter Wechsler-Intelligenztest (2. völlig überarbeitete und erweiterte Auflage). Meisenheim: Hain.

Georgas, J., Weiss, L., van de Vijver, F. & Saklofske, D. (Eds.). (2003). Cultures and Children's Intelligence: A Cross-Cultural Analysis of the WISC-III. New York: Academic Press.

Hardesty, F. P. & Priester, H. J. (1956). Handbuch für den Hamburg-Wechsler-Intelligenztests für Kinder HAWIK (3. Auflage 1966). Bern: Huber.

Kaufman, A. S. (1979). A four-test short form of the WISC-R. Contemporary Educational Psychology, 1, 180-196.

Lienert, G & Raatz, U. (1998). Testaufbau und Testanalyse (6. Auflage). Weinheim: Psychologie Verlags Union.

Nunnally, J. C. & Bernstein, I. H. (1994). Psychometric theory (third edition). New York: McGraw-Hill.

Pawlik, K. (1964). Altersspezifische Kurzformen im Hamburg-Wechsler-Intelligenztest für Kinder (HAWIK). Psychologie und Praxis, 8, 14-26.

Schäuble, R. & Gorlicki, C. (1998). Ist mit einer reduzierten Form des HAWIE-R eine valide Intelligenzeinstufung möglich? Diagnostica, 44, 5-10.

Tabachnick, B. G. & Fidell, L. S. (1996). Using multivariate statistics (3. Auflage). New York: Harper & Row.

Tewes, U. (1983). HAWIK-R. Handbuch und Testanweisung. Bern: Huber.

Tewes, U., Rossmann, P. & Schallberger, U. (1999). HAWIK-III. Übersetzung und Adaptation der WISC-III von David Wechsler. Bern: Hans Huber.

Tewes, U., Rossmann, P. & Schallberger, U. (2002). HAWIK-III. Übersetzung und Adaptation der WISC-III von David Wechsler (3., überarbeitete und ergänzte Auflage). Bern: Hans Huber.

Thompson, B. & Borrello, G. M. (1985). The importance of structure coefficients in regression research. Educational and Psychological Measurement, 45, 203-209.

Wechsler, D. (1949). WISC. New York: Psychological Corporation.

Wechsler, D. (1974). WISC-R. San Antonio, TX: Psychological Corporation.

Wechsler, D. (1991). WISC-III. San Antonio, TX: Psychological Corporation.

Wechsler, D. (2004). WISC-IV. San Antonio, TX: Harcourt Assessment.

Anhang

Tabelle A-1.
Angaben zu den 24 besten Vierer-Kurzformen für den Altersbereich von 6-16 Jahre

Nummer	Standardschätzfehler (IQ) alle	6-7 Jahre	8-12 Jahre	13-16 Jahre	Schätzformel für die Wertpunktsumme
4-1	5.0	5.7	4.8	4.9	15.83 + 2.37*AW + 1.96*AV + 1.92*BO + 2.20*MT
4-2	5.0	5.8	5.0	4.6	19.31 + 1.95*AW + 2.16*WT + 1.78*BO + 2.22*MT
4-3	5.1	6.0	5.1	4.7	17.68 + 1.94*GF + 2.27*WT + 1.87*BO + 2.19*MT
4-4	5.1	5.8	5.0	4.9	16.83 + 1.71*RD + 2.69*WT + 1.82*BO + 2.12*MT
4-5	5.2	5.5	5.2	4.9	18.48 + 2.03*AW + 2.52*WT + 1.87*MT + 1.76*FL
4-6	5.1	5.1	5.0	5.2	10.66 + 2.08*RD + 2.86*WT + 1.78*BE + 2.19*FL
4-7	5.1	5.3	5.1	4.9	11.95 + 1.99*GF + 1.81*RD + 2.43*WT + 2.55*FL
4-8	5.3	5.8	5.2	5.0	17.09 + 2.20*AW + 2.26*WT + 1.46*ZS + 2.39*MT
4-9	5.2	5.2	5.3	5.2	13.66 + 2.25*GF + 2.47*WT + 1.71*BO + 2.22*FL
4-10	5.3	5.9	5.4	4.9	19.19 + 1.99*AW + 2.19*WT + 1.65*BE + 2.26*MT
4-11	5.2	5.2	5.1	5.2	12.78 + 1.99*RD + 2.94*WT + 1.65*BO + 2.15*FL
4-12	5.2	5.8	5.2	5.0	16.11 + 2.11*GF + 2.60*WT + 1.76*MT + 1.93*FL
4-13	5.2	5.5	5.1	5.1	15.21 + 1.86*RD + 3.04*WT + 1.70*MT + 1.87*FL
4-14	5.2	5.6	5.2	5.1	15.63 + 1.87*RD + 2.66*WT + 1.82*BE + 2.09*MT
4-15	5.3	5.9	5.2	5.1	14.41 + 2.35*GF + 1.95*AV + 2.08*BO + 2.20*MT
4-16	5.3	5.9	5.1	5.3	17.89 + 2.32*WT + 1.57*AV + 1.99*BO + 2.37*MT
4-17	5.4	5.2	5.3	5.5	16.66 + 2.05*AW + 2.49*WT + 1.69*BO + 2.13*FL
4-18	5.3	6.0	5.2	5.1	17.59 + 2.87*WT + 1.54*BE + 1.73*BO + 2.14*MT
4-19	5.4	5.5	5.3	5.3	12.74 + 2.19*GF + 2.51*WT + 1.65*BE + 2.36*FL
4-20	5.4	5.9	5.3	5.3	17.77 + 3.21*WT + 1.67*BO + 1.87*MT + 1.51*FL
4-21	5.4	5.3	5.4	5.5	15.39 + 2.03*AW + 2.49*WT + 1.69*BE + 2.24*FL
4-22	5.3	5.4	5.3	5.3	15.15 + 1.69*AW + 1.73*RD + 2.55*WT + 2.51*FL
4-23	5.4	5.7	5.4	5.3	16.83 + 2.51*AW + 1.85*AV + 1.64*BE + 2.31*MT
4-24	5.4	5.8	5.4	5.2	10.50 + 2.42*GF + 2.61*WT + 1.41*ZS + 2.50*FL

Anmerkungen: In die Schätzformeln sind die auf übliche Weise bestimmten Wertpunkte der jeweiligen Untertests einzusetzen. Der Standardschätzfehler ist in IQ-Punkten ausgedrückt und wurde empirisch als Standardabweichung der Differenz zwischen geschätztem IQ und dem Total-IQ berechnet. Die 6 Kurzformen 4-1 bis 4-6 entsprechen den Kurzformen 4-A bis 4-F im Manual (Tewes et al., 2002, Kapitel 7). Die vereinzelten geringfügigen Abweichungen in den Schätzparametern erklären sich durch Druckfehler im Manual, die vom Autor verschuldet sind.

Tabelle A-2.
Angaben zu den 29 besten Fünfer-Kurzformen für den Altersbereich 6-7 Jahre

Nummer	Standard-schätz-fehler (IQ)	Schätzformel für die Wertpunktsumme	Erweiterung der Vierer-Kurzformen:
5-1	4.2	14.41+1.61*RD+2.15*WT+1.63*BE+1.35*BO+1.83*FL	6
5-2	4.3	13.43+1.74*AW+1.96*WT+1.51*BE+1.51*BO+1.94*FL	17,21
5-3	4.3	14.56+1.38*GF+1.67*RD+1.81*WT+1.65*BE+2.03*FL	6,7,19
5-4	4.4	13.06+1.57*GF+2.02*WT+1.54*BE+1.58*BO+1.99*FL	9,19
5-5	4.3	15.65+1.42*AW+1.52*RD+1.82*WT+1.65*BE+2.02*FL	6,21,22
5-6	4.5	17.88+1.80*RD+2.11*WT+1.53*BE+1.15*MT+1.63*FL	6,13,14
5-7	4.5	7.78+1.58*GF+1.83*RD+2.31*AV+1.35*ZS+2.14*MT	
5-8	4.6	8.65+1.74*AW+1.60*RD+2.21*AV+1.36*ZS+2.22*MT	
5-9	4.4	13.37+1.52*GF+1.46*RD+2.05*WT+1.50*BO+2.14*FL	7
5-10	4.5	12.56+1.60*AW+1.41*GF+1.88*WT+1.63*BO+2.21*FL	9,17
5-11	4.6	15.41+1.93*AW+2.07*WT+1.42*BO+1.29*MT+1.75*FL	2,5,17,20
5-12	4.6	17.59+1.85*RD+1.74*WT+1.58*AV+1.64*MT+1.45*FL	13
5-13	4.6	13.55+1.45*GF+1.94*RD+1.99*AV+1.42*BE+1.83*MT	
5-14	4.7	10.35+2.04*RD+2.26*AV+1.49*BE+1.30*ZS+1.89*MT	
5-15	4.7	16.43+1.96*RD+1.47*WT+1.67*AV+1.43*BE+1.84*MT	14
5-16	4.6	17.39+1.62*AW+1.44*RD+1.94*WT+1.48*MT+1.77*FL	5,13,22
5-17	4.5	13.96+1.82*RD+2.11*WT+1.80*BE+0.94*ZS+1.94*FL	6
5-18	4.6	14.67+1.55*AW+1.29*RD+2.08*WT+1.48*BO+2.13*FL	11,17,22
5-19	4.7	17.55+1.91*AW+1.90*WT+1.41*BE+1.28*MT+1.75*FL	5,10,21
5-20	4.6	14.50+1.57*AW+1.74*RD+1.92*AV+1.40*BE+1.92*MT	23
5-21	4.5	16.95+1.89*RD+1.79*WT+1.10*AV+1.65*BE+1.89*FL	6
5-22	4.6	16.83+1.46*GF+1.66*RD+1.98*WT+1.39*MT+1.84*FL	7,12,13
5-23	4.8	7.13+2.19*AW+2.17*AV+1.45*ZS+1.41*BO+2.07*MT	1
5-24	4.6	15.29+1.61*AW+1.34*GF+1.71*WT+1.60*BE+2.21*FL	19,21
5-25	4.7	14.99+1.77*RD+1.62*WT+1.88*AV+1.33*BO+1.92*MT	4,16
5-26	4.7	11.99+1.90*RD+1.46*WT+2.04*AV+1.23*ZS+2.18*MT	
5-27	4.6	7.30+1.92*AW+1.47*GF+2.06*AV+1.53*ZS+2.28*MT	
5-28	4.6	12.05+1.57*GF+1.77*RD+2.24*AV+1.29*BO+1.93*MT	15
5-29	4.7	14.65+1.63*GF+1.86*RD+2.01*AV+1.69*MT+1.33*FL	

Anmerkungen: In die Schätzformeln sind die auf übliche Weise bestimmten Wertpunkte der jeweiligen Untertests einzusetzen. Der Standardschätzfehler ist in IQ-Punkten ausgedrückt und wurde empirisch als Standardabweichung der Differenz zwischen geschätztem IQ und dem Total-IQ berechnet. Die drei Kurzformen 5-1 bis 5-3 entsprechen den Kurzformen 5-A bis 5-C im Manual (Tewes et al., 2002, Kapitel 7). Die vereinzelten geringfügigen Abweichungen in den Schätzparametern erklären sich durch Druckfehler im Manual, die vom Autor verschuldet sind.