

Andreas Schelten

Testbeurteilung und Testerstellung



Franz Steiner Verlag Stuttgart

Andreas Schelten

Testbeurteilung und Testerstellung

Andreas Schelten

Testbeurteilung und Testerstellung

Grundlagen der
Teststatistik und Testtheorie für Pädagogen
und Ausbilder in der Praxis

Zweite, durchgesehene und erweiterte
Auflage



Franz Steiner Verlag Stuttgart 1997

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Schelten, Andreas:

Testbeurteilung und Testerstellung : Grundlagen der Teststatistik und Testtheorie
für Pädagogen und Ausbilder in der Praxis / Andreas Schelten. - 2., durchges.
und erw. Aufl. - Stuttgart : Steiner, 1997

ISBN 3-515-07141-5



ISO 9706

Jede Verwertung des Werkes außerhalb der Grenzen des Urheberrechtsgesetzes ist unzulässig und strafbar. Dies gilt insbesondere für Übersetzung, Nachdruck, Mikroverfilmung oder vergleichbare Verfahren sowie für die Speicherung in Datenverarbeitungsanlagen. © 1997 by Franz Steiner Verlag Wiesbaden GmbH, Sitz Stuttgart. Gedruckt auf alterungsbeständigem Papier. Druck: Rhein Hessische Druckwerkstätte, Alzey.

Printed in Germany

Vorwort

Der vorliegende Band ist aus erziehungswissenschaftlichen Lehrveranstaltungen über Leistungsmessung und Messung von Persönlichkeitsmerkmalen entstanden. Er richtet sich an statistisch nicht Vorgebildete und verfolgt zwei Ziele:

- Es werden jene teststatistischen und testtheoretischen Kenntnisse und Fertigkeiten vermittelt, die zur Entwicklung eines stichhaltigen Leistungstests erforderlich sind.
- Es wird durch eine Erweiterung dieser Kenntnisse ermöglicht, auf dem Markt befindliche Tests (Intelligenztests, Eignungstests, Konzentrationstests, Sozialtests, Entwicklungstests, Einschulungstests oder Schulleistungstests) nach teststatistischen und testtheoretischen Kriterien angemessen beurteilen und einsetzen zu können.

Mit letzterem sind neben Pädagogen auch Ausbilder angesprochen, die auf Tests oder testähnliche Verfahren bei der Einstellung von Auszubildenden zurückgreifen.

Der Band gliedert sich in zwei Teile. Teil A (Kap. 1–7) befaßt sich mit der deskriptiven Statistik, soweit diese für eine Testerstellung und Testbeurteilung unerlässlich ist (Beschreibung von Testrohwerten, Maße der zentralen Tendenz und Variabilität, Perzentile und Prozentränge, Standardwerte und Standardskalen, Testnormen, Korrelation). – Eine Intervall- oder Klassenbildung bei Testrohwersätzen und die damit verbundene Erörterung statistischer Berechnungen wird nicht durchgeführt. Die Klassen- oder Intervallbildung bei Rohwersätzen ist ein Verarbeitungsschritt, der die Konzepte statistischer Begriffe wie Mittelwert, Median und Standardabweichung mehr zuzudecken als offenzulegen scheint. – Die didaktische Gestaltung des ersten Teils richtet sich danach, in welcher Form eine deskriptive Statistik – angemessen verwandt – dem Pädagogen und Ausbilder dienen kann.

Teil B (Kap. 8–13) hebt nach den Vorbereitungen im Teil A direkt auf die teststatistischen und testtheoretischen Konzepte zur Konstruktion und Beurteilung eines Testes ab. Ausgehend von dem allumfassenden Begriff der Validität werden die notwendigen Bedingungen für eine Validität – die Reliabilität und Objektivität – abgehandelt. Es folgt die Erörterung der Aufgabenanalyse, die zur Sicherstellung einer ausreichenden Reliabilität unerlässlich ist. Regeln zur Gestaltung von Testaufgaben, die eine Hilfe bei der Testkonstruktion und -revision geben, und teststatistische Modelle zur Notengebung schließen den zweiten Teil ab.

Den teststatistischen und testtheoretischen Erörterungen liegt hier die klassische, d. h. eine auf die normbezogene Leistungsmessung ausgerich-

tete Testtheorie zugrunde. Es sei ausdrücklich darauf verwiesen, daß mit den hier dargestellten Verfahren sich ein normbezogener Leistungstest erstellen läßt (Vergleich mit Personengruppen) und nicht ein kriteriumsbezogener Leistungstest (Vergleich auf Lernziel erreicht/nicht erreicht). Ferner ist in erster Linie nicht Gegenstand des Bandes, die Problematik einer Leistungsmessung aufzuzeigen. Es werden hier die teststatistischen Techniken und Konzepte einer Leistungsmessung vermittelt, um von dorthier die Problematik einer Leistungsmessung verständlich werden zu lassen.

Gießen, im Januar 1980

Andreas Schelten

Vorwort zur zweiten Auflage

Die Grundlagen der Testbeurteilung und Testerstellung, in erster Auflage 1980 bei Quelle und Meyer in Heidelberg als Uni-Taschenbuch (UTB 995) herausgegeben, erscheinen nunmehr in zweiter durchgesehener und erweiterter Auflage beim Franz Steiner Verlag in Stuttgart unter dem kurzen Titel Testbeurteilung und Testerstellung. Für die zweite Auflage sind nach Bereinigung einiger Druckfehler die Kapitel 1 bis 13 unverändert erhalten geblieben. Ergänzend sind neu die Kapitel 14 und 15 hinzugekommen.

In Kapitel 14 hat Thomas Heinrich, studentische Hilfskraft am Lehrstuhl für Pädagogik der Technischen Universität München und Studierender für das Höhere Lehramt an beruflichen Schulen, zusammen mit dem Verfasser eine Aktualisierung statistischer Berechnungen mit moderner Software vorgenommen. Dabei wurde das weit verbreitete Standard-Tabellenkalkulationsprogramm Microsoft EXCEL für Windows herangezogen. Die Berechnungen beziehen sich auf die Beispiele in den Kapiteln 1 bis 7. Die Rechenoperationen werden in einfacher Weise dargestellt und ihre Erarbeitung in EXCEL kurz beschrieben.

Für eine spezielle und weiterführende Statistiksoftware sei auf das Programmsystem SPSS für Windows verwiesen. Ergänzend dazu empfiehlt sich das Buch von Joerg M. Diehl und Thomas Staufenberg: Statistik mit SPSS für Windows, Version 6.1, Eschborn: Klotz 1996. Dieses Buch ist eng auf das Lehrbuch von Joerg M. Diehl und Heinz U. Kohr über Deskriptive Statistik im gleichen Verlag abgestimmt, das ebenfalls dem Leser empfohlen werden kann, der tiefer in die deskriptive Statistik einsteigen möchte als sie hier in den Kapiteln 1 bis 7 niedergelegt ist.

In Kapitel 15 werden neuere Entwicklungen für eine Testbeurteilung und Testerstellung, wie sie hier dargestellt sind, angesprochen. Es zeigt sich, daß die Grundlagen der Testbeurteilung und Testerstellung nach der klassischen Testtheorie weiterhin Gültigkeit haben. Die Ausführungen werden mit Hinweisen auf ausgewählte Lehrbücher zur pädagogischen Diagnostik beschlossen.

München, im April 1997

Andreas Schelten

Inhalt

Vorworte	5
Einleitung	10
Teil A	
1. Beschreibung von Rohwertsätzen	14
1.1 Urliste und Rangfolgenverteilung	14
1.2 Primäre und sekundäre Häufigkeitsverteilung	17
2. Maße der zentralen Tendenz	20
2.1 Median (Md), Modalwert (Mo), Arithmetischer Mittelwert ($\bar{x}$)	20
2.2 Symmetrische und schiefe Verteilungen	28
2.3 Die Auswahl eines Maßes zentraler Tendenz	29
3. Maße der Variabilität	32
3.1 Variationsweite	32
3.2 Quartilabweichung	33
3.3 Standardabweichung	34
4. Perzentile und Prozenträge	43
5. Standardnormalverteilung (z-Verteilung)	47
5.1 Tabelle	49
5.2 Standardwerte z	53
5.3 Transformation der Standardwerte z in Standardwertskalen (IQ, C, T und Z)	58
5.4 Exkurs: Der ältere IQ-Begriff	60
5.5 Umrechnung von Standardwertskalen	61
5.6 Exkurs: Nicht lineare Transformationen	65
5.7 Stanine, Stasix und Sten	67
6. Normen	73
6.1 Tabellarische Normen	73
6.2 Grafische Normen	79
6.3 Normierung (Standardisierung) eines informellen Testes	81
7. Korrelation	83
7.1 Das Konzept der Korrelation	83
7.2 Berechnungsverfahren des Korrelationskoeffizienten	86
7.3 Interpretation des Korrelationskoeffizienten	91
7.4 Korrelation und Kausalität	93
Teil B	
8. Validität	98
8.1 Der Begriff der Validität	98
8.2 Inhaltsvalidität	98
8.3 Kriterienbezogene Validität	99

8.4	Konstruktvalidität	100
8.5	Zusammenfassung	101
9.	Reliabilität	102
9.1	Der Begriff der Reliabilität	102
9.2	Der Standardmeßfehler	103
9.3	Die Bestimmung des Reliabilitätskoeffizienten	108
9.31	Testhalbierungs-Reliabilität	110
9.32	Konsistenzanalyse	112
9.33	Paralleltest-Reliabilität	115
9.34	Retest-Reliabilität	115
9.35	Zusammenfassung	116
9.4	Möglichkeiten der Reliabilitätsverbesserung	117
9.5	Der Standardmeßfehler und sein Vertrauensintervall	118
9.6	Exkurs: Die Grundaussagen (Axiome) der klassischen Testtheorie	122
10.	Objektivität	124
10.1	Der Begriff der Testobjektivität	124
10.2	Durchführungsobjektivität	125
10.3	Auswertungsobjektivität	125
10.4	Interpretationsobjektivität	126
11.	Aufgabenanalyse	128
11.1	Schwierigkeitsgrad und Trennschärfe	128
11.2	Vereinfachtes Verfahren zur Trennschärfe- und Schwierigkeitsbestimmung (Diskriminationsindex D)	132
11.3	Zusammenhang zwischen Trennschärfe und Schwierigkeit: Aufgabenselektion	134
11.4	Distraktorenanalyse	136
11.5	Aufgabenanalyse für den Beispielttest: Vitamine und Mineral- stoffe	138
12.	Testaufgaben und ihre Konstruktion	144
12.1	Richtig-Falsch-Aufgaben	144
12.2	Mehrfachwahlaufgaben	146
12.3	Zuordnungsaufgaben	149
12.4	Umordnungsaufgaben	151
12.5	Ergänzungsaufgaben	152
12.6	Kurzantwortaufgaben	155
13.	Notengebung	157
13.1	Umsetzung der Rohwerte in Noten	157
13.2	Problematik eines Zensierungsmodelles nach der Normalver- teilung	160

14. Aktualisierung statistischer Berechnungen mit moderner Software	
Thomas Heinrich	163
14.1 Urliste, Rangfolgenverteilung und Primäre Häufigkeitsverteilung (f)	163
14.2 Median (Md), Modalwert (Mo), Arithmetischer Mittelwert ($\bar{x}$)	168
14.3 Variationsweite (V), Quartilabweichung (Q), Standardabweichung (s), Varianz (s^2)	172
14.4 Perzentile und Prozentränge	173
14.5 Standardwerte z, Transformation der Standardwerte z	174
14.6 Berechnung des Korrelationskoeffizienten r	176
15. Neuere Entwicklungen	179
Anmerkungen	182
Literaturverzeichnis	185
Anhang: Lösungen zum Test: Vitamine und Mineralstoffe	186
Sachregister	187

Einleitung

Wurden die schriftlichen Arbeiten, die Sie in der Schule geschrieben haben, noch „Klassenarbeiten“ genannt, oder wurden sie gelegentlich schon als „Tests“ bezeichnet? Ob „Klassenarbeit“ oder „Test“, in beiden Fällen werden wohl einzelne Teile mit Punkten versehen worden sein. Dies geschah zur Objektivierung der Notengebung.

Nicht die Punktvergabe allein kennzeichnet eine solche Objektivierung, sondern erst die statistisch-rechnerische Absicherung der Notenermittlung: Eine Leistungserfassung sollte nach teststatistischen und testtheoretischen Kriterien geplant, durchgeführt und ausgewertet werden. Auch in den modernen, vom Lehrer selbst erstellten „informellen Tests“ sollte die Leistungserfassung hinsichtlich dieser Kriterien objektiviert sein.

Die teststatistischen und testtheoretischen Kenntnisse und Verfahrensweisen, die zur Konstruktion eines Testes notwendig sind, sollen Ihnen im folgenden vermittelt werden. Zur Veranschaulichung ist ein informeller Test auf den folgenden Seiten abgedruckt¹, an dem die einzelnen Schritte von Fall zu Fall demonstriert werden.

Eine Unterrichtsreihe über „Vitamine und Mineralstoffe“ in einer Berufsfachschule des medizinisch-technischen und krankenschwägerischen Berufsfeldes ging dem Test voraus. 30 Schüler und Schülerinnen bearbeiteten den Test. Es ist die nicht revidierte, erste Fassung des Testes, die Sie hier vor sich haben. Die teststatistische Bearbeitung der Ergebnisse dieses Testes wird später Konstruktionsfehler aufdecken. – Der Test umfaßt 20 Aufgaben. Es wird teils die freie, teils die gebundene Beantwortung der Fragen gefordert (vgl. Kap. 12). Pro Aufgabe wird ein Punkt vergeben. Die Lösungsantworten sind im Anhang aufgeführt.

Name:

Klasse:

Datum:

Test: Vitamine und Mineralstoffe

Arbeitsanweisung: Bei vorgegebenen Antworten die zutreffenden (gegebenenfalls mehrere) ankreuzen! Freie Stellen mit kurzen Antworten füllen und wo gefordert, eine Begründung angeben!

1. Zu welchen Stoffen zählen wir die Vitamine? Zu den
a) Baustoffen ☐
b) Wirkstoffen ☐
c) Brennstoffen ☐
2. Zu welchen Stoffen zählen wir die Mineralstoffe? Zu den
a) Brennstoffen ☐
b) Wirkstoffen ☐
c) Baustoffen ☐
3. In welche beiden Gruppen unterscheiden wir die Vitamine?

4. Nach der Bedarfsmenge unterscheiden wir 2 Gruppen von Mineralstoffen:

5. Welche beiden Gruppen von Nahrungsmitteln gelten als besonders vitaminreich?

6. Wo finden wir den größten Anteil fettlöslicher Vitamine?
a) in der Kartoffel ☐
b) im Getreide ☐
c) im Ei ☐
7. Durch vitaminarme bzw. vitaminfreie Kost können _____

8. Wodurch wird die Wirksamkeit der Vitamine beeinträchtigt?
a) Wasser ☐ | c) Luft ☐ | e) Trockenheit ☐
b) Kälte ☐ | d) Hitze ☐ | f) Licht ☐
9. Wie lange kann man Nahrungsmittel mit einem hohen Anteil fettlöslicher Vitamine wässern, ohne daß der Vitamingehalt beeinträchtigt wird? Begründe Deine Antwort kurz!
a) nur kurz _____ ☐
b) längere Zeit _____ ☐
c) unbegrenzt _____ ☐
10. Welchen chemischen Stoff bezeichnet man mit Vitamin C?

11. In welcher Größenordnung liegt der tägliche Bedarf in Vitaminen?
(Keine Zahlenangabe)

12. Warum gehören die Vitamine und Mineralstoffe für den Körper zu den essentiellen Stoffen? _____

13. Gib 3 Dir bekannte Mengenelemente an!
1. _____ 2. _____ 3. _____
14. Gib 3 Dir bekannte Spurenelemente an!
1. _____ 2. _____ 3. _____
15. Welches Nahrungsmittel ist besonders reich an Mineralstoffen (vor allem an Ca)? _____
16. Ordne die beiden Gruppen ihrer hauptsächlichlichen Funktion zu! (Die Zuordnung geht nicht auf!)
- | | |
|-------------------|-------------------|
| 1. Mengenelemente | ___ a) Brennstoff |
| 2. Spurenelemente | ___ b) Wirkstoff |
| | ___ c) Baustoff |
17. Sind Mineralstoffe löslich?
Ja ☐ Nein ☐
Wenn ja, worin? _____
18. Wie unterscheidet sich der Bedarf an Mineralstoffen bei Kindern und Erwachsenen? _____

19. Warum sollte man ungeschälten Reis geschältem Reis vorziehen? _____

20. Wie nennt man die Vorstufe von Vitaminen? _____

Teil A

1. Beschreibung von Rohwertsätzen

1.1 Urliste und Rangfolgenverteilung

Bei der statistischen Betrachtung von Tests wird nicht von Punkten, sondern von Rohwerten gesprochen. Man spricht von rohen Werten, weil später die Rohwerte in aussagefähigere, testspezifische Werte transformiert werden (siehe unten Standardwerte). – Entsprechend wird von Rohwertträgern gesprochen. Rohwertträger sind in einem Schulleistungstest die Schüler, die den Test bearbeiten.

In einem ersten Auswertungsschritt werden die Rohwerte den Rohwertträgern zugeordnet. Zum Beispiel (s. u.) erreichte der Rohwertträger Michael E. in dem Test über Vitamine und Mineralstoffe 8 Rohwerte. Er hatte 8 von 20 Aufgaben richtig gelöst.

Rohwertträger Schüler	Rohwert (RW)
Michael E.	8
Beatrix H.	15
Karin E.	13
Friedel S.	17
usw.	

Werden in dieser Form ohne jegliche Ordnungsgesichtspunkte die Ergebnisse eines Testes aufgelistet, erhält man eine Urliste.

Definition 1:

Die Anordnung der Rohwerte, bei der lediglich Rohwert und Rohwertträger einander zugeordnet sind, sonstige Ordnungsgesichtspunkte aber fehlen, wird *Urliste* genannt.

Die Urliste kann nur einen ungefähren Einblick in die Leistung eines Schülers bieten. Es ist nur schwer zu ersehen, ob die Schülerin Karin E. im Vergleich zu ihren Mitschülern eine mittlere oder hohe Leistung erzielte. Ist Michael E. im Vergleich zu seinen Mitschülern ein ausgesprochen schlechter Schüler? Oder tendiert seine Leistung etwa schon zu der von den mittleren Schülern?

Die Rohwerte müssen zur Beantwortung dieser Fragen in eine Rangfolge gebracht werden. Dabei werden die Rohwerte der Höhe nach geordnet. Der höchste Wert erhält den Rang 1. Tabelle 1 gibt die Rangfolgen-Verteilung für den Test Vitamine und Mineralstoffe wieder.

Rohwertträger			Rohwertträger		
Schüler	Rohwert	Rang	Schüler	Rohwert	Rang
Christa H.	18	1	Erika B.	11	19
Friedel S.	17	2	Marion B.	11	19
Ilona W.	16	3,5	Jutta H.	11	19
Elvira G.	16	3,5	Cornelia F.	11	19
Beatrix H.	15	5,5	Siegrid L.	11	19
Carla G.	15	5,5	Sonja S.	10	23
Christine W.	14	8	Thomas B.	10	23
Helga W.	14	8	Ingeborg E.	10	23
Christa W.	14	8	Gudrun P.	8	26,5
Angela R.	13	12	Michael E.	8	26,5
Christel G.	13	12	Bärbel W.	8	26,5
Regina N.	13	12	Ilse N.	8	26,5
Britta A.	13	12	Hilde M.	7	29
Karin E.	13	12	Marlies G.	6	30
Astrid P.	12	15,5			
Ute B.	12	15,5			

Tab. 1:

Rangfolgen-Verteilung der Rohwerte im Test Vitamine und Mineralstoffe, N = 30 Schüler, maximal erreichbarer Rohwert 20.

Definition 2:

Der *Rang* macht eine Aussage über die relative Position eines Rohwertträgers (z. B. eines Schülers) innerhalb eines getesteten Personenkreises (z. B. einer Schulklasse). Der Rohwertträger mit dem höchsten Rohwert erhält dabei den Rang 1.

Ilona W. und Elvira G. erhalten den gleichen Rang. Beide erzielten im Test den Rohwert 16. Werden die Rohwerte in Tabelle 1 von oben nach unten durchnumeriert, steht der Rohwert 16 von Ilona W. an 3. Stelle, der von Elvira G. an 4. Stelle. Der Durchschnitt aus Rang 3 und 4 ist 3,5; rechnerisch: $(3+4)/2 = 3,5$. Die Rohwerte der Schülerinnen Christine W., Helga W. und Christa W. stehen an 7., 8. und 9. Stelle. Der Durchschnitt ist der Rang 8; rechnerisch: $(7+8+9)/3 = 8$.

Übung 1

Numerieren Sie bitte die Rohwerte in Tabelle 1 von oben nach unten durch und prüfen Sie nach, ob die Schüler Gudrun P., Michael E., Bärbel W. und Ilse N. zu Recht den Rang 26,5 erhalten.

Aus der Rangfolgenverteilung läßt sich ablesen, daß die Schülerin Karin E. im Vergleich zu ihren Mitschülern eine durchschnittliche Leistung einnimmt. Mit einem Rohwert von 13 aus 20 möglichen Rohwerten steht sie in der Schülerrangreihe an 12. Stelle. Michael E. steht mit seiner Leistung in der Rangreihe ziemlich am Ende. Von 30 Schülern hat er den Rang 26,5.

Im folgenden finden Sie die vollständige Grundausswertung des Testes Vitamine und Mineralstoffe vor. Die Daten sind bereits nach der Leistung geordnet, von der höchsten bis zur niedrigsten. Rechts, neben die Spalte „Punktwert des Schülers“, kann noch der Rang des Schülers eingetragen werden, wie es hier bereits in Tabelle 1 geschehen ist.

Aufgabe																					Punktwert des Schülers (Summe)
Schüler	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
Christa H.	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0	+	+	+	/	19
Friedel S.	0	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0	+	+	+	/	17
Ilona W.	0	+	+	+	+	+	+	+	+	+	+	+	+	+	0	0	+	+	+	/	16
Elvira G.	+	+	+	+	+	0	+	0	0	+	+	+	+	+	0	+	+	+	+	+	16
Carla G.	+	0	+	+	+	0	+	+	0	+	+	+	+	+	+	0	+	+	+	/	15
Beatrix H.	+	0	+	+	0	+	+	+	+	+	+	+	+	+	/	+	0	+	+	/	15
Christine W.	0	+	+	+	+	0	+	0	+	+	+	+	+	+	0	+	0	+	+	/	14
Christa W.	+	+	+	+	0	+	0	/	+	0	+	+	+	+	0	+	+	+	+	0	14
Helga W.	+	+	0	+	0	+	+	0	0	0	+	+	+	+	0	+	+	+	0	+	14
Angela K.	+	0	+	+	0	0	+	+	+	+	0	+	+	0	+	/	+	+	/	13	
Christel G.	+	0	+	+	0	0	+	+	0	+	+	+	+	+	0	0	+	+	+	/	13
Regina M.	0	0	+	+	+	0	+	+	0	+	+	+	+	+	0	0	+	+	+	/	13
Britta A.	+	+	+	+	0	+	+	+	+	0	0	+	+	0	0	0	0	+	+	+	13
Karin E.	0	+	+	+	+	0	0	+	+	0	0	+	+	+	0	+	0	+	+	+	13
Astrid P.	+	0	+	+	+	+	0	+	0	+	/	+	+	0	0	0	0	+	+	/	12
	10	9	14	15	9	9	11	14	8	12	11	13	15	14	6	4	9	15	14	4	

Aufgabe																					Punktwert des
Schüler	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Schülers (Summe)
Ute B.	+	0	+	+	/	0	+	+	0	+	+	/	+	+	0	+	0	+	+	/	12
Erika D.	+	0	0	+	+	0	+	+	+	+	+	0	/	0	0	0	+	+	+	/	11
Marion B.	0	+	+	0	+	+	+	0	/	0	+	+	+	0	0	0	+	0	+	+	11
Jutta H.	0	0	+	+	0	+	+	/	0	/	+	+	+	+	0	0	+	/	+	+	11
Cornelia F.	0	0	+	+	0	0	+	+	0	0	/	+	+	+	+	0	0	+	+	+	11
Siegrid L.	+	0	+	+	0	+	+	+	0	/	+	0	+	+	+	0	0	+	0	/	11
Senja S.	+	0	+	+	/	+	/	+	+	+	+	/	+	0	0	0	/	/	+	/	10
Thomas B.	0	0	/	+	+	0	+	0	0	+	+	+	0	/	+	0	+	+	+	/	10
Ingeborg E.	0	+	+	+	+	0	+	0	/	0	+	0	0	0	0	0	0	+	+	+	10
Gudrun P.	+	0	/	+	+	+	/	0	0	+	+	/	/	/	/	0	0	+	+	/	8
Michael E.	0	0	+	+	+	0	0	+	0	+	0	/	+	+	/	0	0	+	/	/	8
Bärbel W.	0	0	+	+	0	+	+	+	0	0	/	+	0	0	/	0	0	+	+	/	8
Ilse M.	+	0	+	+	+	0	0	0	0	/	+	0	0	/	0	0	+	+	+	+	8
Hilde M.	0	0	+	+	0	+	+	0	0	0	+	+	0	/	0	0	0	+	0	/	7
Marlis G.	0	0	+	/	0	+	0	+	0	0	0	+	+	0	0	0	0	0	+	+	6
	6	2	12	14	7	8	9	11	2	6	7	10	8	6	3	2	0	13	10	6	

Zeichenerklärung: + Aufgabe richtig
 0 " falsch
 / " nicht beantwortet

Tab. 2:
Grundausswertung des Tests Vitamine und Mineralstoffe

Die Bestimmung der Ränge hat einen großen Nachteil. Ein Vergleich mit anderen Klassen, deren Umfang beträchtlich größer oder kleiner sein kann, läßt sich nicht anstellen. Ein Rang von 20 in einer Klasse von 25 bedeutet etwas anderes, als z.B. ein Rang von 20 in einer Klasse von 45 Schülern.

1.2 Primäre und sekundäre Häufigkeitsverteilung

Die Urliste läßt sich weiter vereinfachen, wenn die Rohwerte in einer Häufigkeitsverteilung angeordnet werden. Tabelle 3 zeigt für den hier angeführten Beispielttest die einfachste Art einer solchen Verteilung, die „Primäre Häufigkeitsverteilung“. In Tabelle 3 sind die Rohwertträger nicht mehr aufgeführt. Sie finden hier nur noch die erzielten Rohwerte mit den Besetzungshäufigkeiten vor. – Klein f steht für Frequenz.

Rohwert (RW)	Häufigkeit (f)
18	1
17	1
16	2
15	2
14	3
13	5
12	2
11	5
10	3
8	4
7	1
6	1
Summe	30

Tab. 3:

Primäre Häufigkeitsverteilung der erreichten Rohwerte im Test Vitamine und Mineralstoffe, maximal erreichbarer Rohwert: 20.

Von einer „Sekundären Häufigkeitsverteilung“ spricht man, wenn die Rohwerte in Rohwertklassen (Intervalle), z. B. RW = 18–16, 15–13 usw., zusammengefaßt und die Häufigkeiten der Rohwerte in diesen Intervallen bestimmt werden. – Im folgenden soll aber nur die Primäre Häufigkeitsverteilung von Interesse bleiben.

Aus der Tabelle 3 ist zu ersehen, daß zwischen den Rohwerten 6 und 18 jeder mögliche Rohwert besetzt ist, bis auf den Rohwert 9. Dies muß nicht unbedingt der Fall sein, wenn Sie mit einem Test arbeiten, bei dem z. B. 50 oder mehr Rohwerte erreicht werden können. Bei der Erstellung der Primären Häufigkeitsverteilung ist es zweckmäßig, jeden möglichen Rohwert innerhalb der Rohwertgrenzen, die erreicht wurden, aufzulisten,

auch wenn seine Besetzungshäufigkeit Null ist. Die rechte Seite der Tabelle 4 auf Seite 18 zeigt Ihnen das Vorgehen. 36 Schüler eines ersten Schuljahres wurden einem „Lesereifetest“ (Reading Readiness Test) unterzogen. 128 Rohwerte waren maximal zu erreichen. Das Auflisten auch der nicht besetzten Rohwerte hat folgenden Vorteil: „Löcher“ oder Sprünge in der Verteilung sind erkennbar. Bei Gruppierungsentscheidungen ist der teilende Rohwert dort zu suchen, wo der niedrigste Rohwert in der einen Gruppe und der höchste in der anderen Gruppe liegt. Hierzu ein Beispiel²: Nehmen Sie an, Sie sollten aufgrund der Ergebnisse dieses Lesereifetestes „homogene“ Gruppen zur Binnendifferenzierung eines Unterrichtes bilden. Es würde z. B. viel besser sein, den minimalen Rohwert für die leistungsstärkste Gruppe bei Rohwert 89 zu setzen als bei Rohwert 90. – Nehmen Sie andererseits an, die Primäre Häufigkeitsver-

Rohwert	Rang	Rohwert	Häufigk.	Rohwert	Häufigk.
112	1	112	1	76	0
109	2	111	0	75	3
106	3	110	0	74	1
105	4	109	1	73	0
104	5	108	0	72	1
100	6	107	0	71	1
97	7	106	1	70	1
95	8,5	105	1	69	1
95	8,5	104	1	68	1
93	10	103	0	67	0
91	11	102	0	66	1
90	12	101	0	65	0
89	13	100	1	64	0
84	14,5	99	0	63	0
84	14,5	98	0	62	1
83	16	97	1	61	0
82	17	96	0	60	0
81	18	95	2	59	1
80	19	94	0	58	1
78	20	93	1	57	0
75	22	92	0	56	0
75	22	91	1	55	0
75	22	90	1	54	0
74	24	89	1	53	0
72	25	88	0	52	0
71	26	87	0	51	1
70	27	86	0	50	0
69	28	85	0	49	0
68	29	84	2	48	0
66	30	83	1	47	1
62	31	82	1	46	0
59	32	81	1	45	0
58	33	80	1	44	1
51	34	79	0		
47	35	78	1		
44	36	77	0		

N = 36

Tab. 4:

Rangfolgenverteilung und Primäre Häufigkeitsverteilung der Rohwerte in einem „Lesereifetest“ (nach J. C. Stanley, K. D. Hopkins, Educational and Psychological Measurement and Evaluation, Englewood Cliffs, New Jersey 1972, S.22).

teilung in Tabelle 4 würde das Ergebnis eines Schulleistungstestes darstellen. Sie wollten die Note „Eins“ eingrenzen. Es würde sicher vertretbarer sein, den Rohwert 104 als den Minimalrohwert für eine Eins festzusetzen als die Rohwerte 105 oder 106. Bilden Sie nur eine Rangfolgenverteilung der Testwerte, setzen Sie sich der Gefahr aus, diese sinnvollen Gruppierungsentscheidungen nicht zu finden. Die Rangfolgenverteilung ist zur Verdeutlichung in Tabelle 4 mit aufgenommen.

Definition 3:

In der *Primären Häufigkeitsverteilung* werden die Rohwerte ohne Klassen oder Intervallbildung zusammen mit ihrer Besetzungshäufigkeit aufgelistet. Jeder mögliche Rohwert zwischen kleinstem und größtem tatsächlich auftretenden Wert wird mit seiner Häufigkeit aufgelistet.

Die Primäre Häufigkeitsverteilung verleitet nicht dazu, die Lücken in einer Rohwertverteilung zu übersehen. Bei einer Sekundären Häufigkeitsverteilung kann dies leicht geschehen. Durch geschickte Wahl der Rohwertklassen kann aus einer lückenhaften Verteilung künstlich eine Verteilung entstehen, die keine Lücken aufweist.

Bei der Testentwicklung sollte immer von der Primären Häufigkeitsverteilung ausgegangen werden. Eine Sekundäre Häufigkeitsverteilung sollte nur erstellt werden, wenn zuvor auch die Primäre erstellt wurde. Bei der Sekundären Häufigkeitsverteilung gehen Informationen über das Verteilungsbild und damit über die Güte des Testes verloren.

Rangfolgen- und Häufigkeitsverteilung verdichten in einem ersten Schritt die Information, die in einer Urliste von Rohwerten enthalten ist. Der zweite Schritt besteht darin, die Information eines Satzes von Rohwerten auf statistische Kennwerte zu reduzieren, die in *einer* Zahl etwas über die Verteilung der Rohwerte aussagen. Maße der zentralen Tendenz und Maße der Streuung leisten diese Informationsreduktion. Hierbei gehen naturgemäß Informationen verloren, wie später noch gezeigt wird.

2. Maße der zentralen Tendenz

2.1 Median (Md), Modalwert (Mo), Arithmetischer Mittelwert ($\bar{x}$)

Mit einem Maß der zentralen Tendenz läßt sich das Leistungsniveau der Ergebnisse eines Testes beschreiben. Durch einen Zentralwert soll mitgeteilt werden, wo der „Rohwertepulk“ in einem Satz von Rohwerten liegt. Sind in dem Beispielttest Vitamine und Mineralstoffe maximal 20 Rohwerte zu erreichen, um welchen Rohwert zentrieren sich die Rohwerte? Ist es der Rohwert 11, 12 oder 13? Im folgenden wird auf dieses Beispiel noch eingegangen. Erklärt seien vorab die verschiedenen Maße der zentralen Tendenz.

Median

Definition 4:

Der *Median* ist der Rohwert, der die in eine Rangfolge gebrachten Rohwerte so in zwei Hälften teilt, daß die eine Hälfte der Werte größer, die andere Hälfte kleiner als der Median ist. Schreibweise: „ Md “.

Die Bestimmung des Medians ist unproblematisch, solange Rohwerte nicht mehrmals auftreten, insbesondere beim oder nahe des Medians. Trifft dies zu, lassen sich zwei Fälle unterscheiden:

- (1) Ungerade Zahl von Werten, kein Rohwert tritt doppelt auf: Der Median ist der mittlere der in eine Rangfolge gebrachten Werte:

$$10, 11, 13, \textcircled{15}, 16, 18, 19. Md = 15.$$

- (2) Gerade Zahl von Rohwerten, kein Wert tritt doppelt auf: Der Median ist das Mittel der zwei zentralen Werte der in eine Rangfolge gebrachten Rohwerte:

$$3, 5, 6, \textcircled{7}, \textcircled{8}, 10, 12, 13. Md = (7 + 8)/2 = 7,5.$$

Für den Rohwertsatz in Tabelle 4, Seite 18 läßt sich der Median unproblematisch bestimmen. $N = 36$ Schüler wurden auf ihre Lesereife getestet. Die Anzahl der Rohwerte ist gerade. Der Median ist nach Fall (2) das Mittel der zwei zentralen Werte. Zum Auffinden der zwei zentralen Werte wird von der Gesamtzahl der getesteten Personen in folgender Weise ausgegangen: 36 minus 2 ist gleich 34. Die Hälfte davon ist 17. Links und rechts der zwei zentralen Werte müssen 17 Rohwerte liegen. Zählen Sie in Tabelle 4, Seite 18, in der Spalte Häufigkeit, bei Rohwert 112 beginnend, von oben nach unten, 17 Rohwertbesetzungen ab. Sie gelangen bis einschließlich Rohwert 82. Zählen Sie dann von unten nach oben, beginnend

mit Rohwert 44, gleichfalls 17 Rohwertbesetzungen ab. Sie gelangen bis einschließlich Rohwert 78. Die zwei zentralen Werte sind nun die Rohwerte 81 und 80, zumal der Rohwert 79 nicht besetzt ist. Der Median ist das Mittel beider Werte, nämlich $Md = (81 + 80)/2 = 80,5$. Links und rechts dieses Wertes liegen 50% der getesteten Schüler.

Im Testbeispiel Vitamine und Mineralstoffe treten Rohwerte beim Median oder nahe des Medians mehrfach auf. Mit einfachen Überlegungen wie oben ist der Median nicht mehr zu bestimmen. Es läßt sich in diesem Fall der Median über eine komplizierte Formel berechnen³. In Tabelle 5, Seite 22, wird statt dessen die Berechnung des Medians in 6 Schritte zergliedert vorgenommen. Bevor Sie die Schritte nachvollziehen, bedarf es noch einiger Erklärungen.

In der Spalte „ cum_f “, Tabelle 5, S.22, sind von unten nach oben die Häufigkeiten aufsummiert worden (Cumulierte Häufigkeit). Den Rohwert 6 erreicht 1 Schüler, den Rohwert 7 gleichfalls 1 Schüler. Rohwert 6 oder 7 erreichen dann 2 Schüler. In Spalte cum_f steht daher auf Höhe des Rohwertes 7 eine 2. Den Rohwert 8 erreichen 4 Schüler. 2 (Spalte cum_f , Besetzungshäufigkeit der Rohwerte 6 oder 7) plus 4 (Spalte Häufigkeit f , Besetzungshäufigkeit des Rohwertes 8) sind gleich 6. Das heißt, 6 Schüler erreichen 8 oder weniger Rohwerte. Aus der Spalte cum_f können Sie ablesen, wieviele Schüler bei oder unter einem bestimmten Rohwert liegen. Z. B. erreichen 26 Schüler 15 oder weniger als 15 Rohwerte. Oder anders ausgedrückt: 26 Schüler erreichen höchstens 15 Rohwerte.

Nach der testtheoretischen Annahme wird die Leistung, die ein Test mißt, als kontinuierlich angesehen. Wird pro Aufgabe ein Rohwert vergeben, heißt es, die Leistungsvariable wird „auf den Punkt genau“ gemessen. Obwohl die vergebenen Punktwerte keine Zwischenwerte zulassen, ist doch die gemessene Leistung, z. B. in den „Kenntnissen über Vitamine und Mineralstoffe“, kontinuierlich. Das heißt, die Leistung liegt nicht genau auf einem gegebenen Punktwert, sondern liegt auch in dem Raum zwischen zwei aufeinander folgenden Punktwerten: Erreicht ein Schüler z. B. 16 Rohwerte, so liegt der Rohwert im „Intervall“ von 15,5 bis 16,5. Korrekter ausgedrückt heißt es z. B. in Tabelle 5: Nicht 26 Schüler erreichen höchstens 15 Rohwerte, sondern 26 Schüler erreichen höchstens $15 + 0,5 = 15,5$ Rohwerte (obere „Intervall- oder Klassengrenze“).

Definition 5:

Die *Cumulierte* (stetig aufsummierte) *Häufigkeit*, cum_f , gibt die Zahl der Rohwertträger (z.B. Schüler) wieder, die unter einem gegebenen Rohwert + 0,5 liegen (Gegebener Rohwert + 0,5 = obere Intervall- oder Klassengrenze).

Die Berechnung des Medians ersehen Sie nun aus den 6 Schritten, die neben und unter Tabelle 5 aufgeführt sind. Der Median beträgt 12,0. Das heißt, daß bei dem Test Vitamine und Mineralstoffe der Rohwert 12 die in eine Rangfolge gebrachten Werte in zwei Hälften teilt. 50% der getesteten Berufsfachschüler erreichen Rohwerte, die unter 12 liegen, und 50% erzielen Werte, die über 12 liegen (maximal waren 20 Rohwerte zu erreichen).

Rohwert (RW)	Häufigkeit (f)	cum _f	Schritte zur Bestimmung des Medians
18	1	30	1. Bestimmen Sie $\frac{1}{2} \cdot N$. Dies führt zu der gesuchten cumulierten Häufigkeit, dessen Rohwertentsprechung den Median ergibt. <u>Im Beispiel:</u> $\frac{1}{2} \cdot N = \frac{1}{2} \cdot 30 = 15.$ 2. Suchen Sie in der Spalte cum_f das Rohwertintervall auf, das die in Schritt 1 bestimmte cumulierte Häufigkeit enthält. <u>Im Beispiel:</u> siehe Kasten (14. Häufigkeit (cum_f = 14) + Häufigkeit f = 1 = 15. Häufigkeit. 15. Häufigkeit + Häufigkeit f = 1 = 16. Häufigkeit (cum_f = 16)).
17	1	29	
16	2	28	
15	2	26	
14	3	24	
13	5	21	
12,5	12	16	
11,5	11	5	
	10	3	
	9	0	
	8	4	
	7	1	
	6	1	
Summe	N = 30		

3. Bestimmen Sie

- a) den Abstand zwischen der in Schritt 1 bestimmten cumulierten Häufigkeit und der nächst niedrig vorkommenden cumulierten Häufigkeit.
- b) die Differenz der zwei vorkommenden cumulierten Häufigkeiten, zwischen der die in Schritt 1 bestimmte cumulierte Häufigkeit liegt.

Im Beispiel:

- a) Die Differenz zwischen der 15. Häufigkeit und der 14. Häufigkeit (cum_f = 14) ist: 15 - 14 = 1 (a)
- b) Die Differenz zwischen der 16. Häufigkeit (cum_f = 16) und der 14. Häufigkeit (cum_f = 14) ist: 16 - 14 = 2 (b)

4. Bestimmen Sie den Quotienten der Größen a und b, die in Schritt 3 bestimmt worden sind. Der Quotient $\frac{a}{b}$ ergibt den Anteil im Rohwertintervall von Schritt 2, der zur Entsprechung der gesuchten cumulierten Häufigkeit benötigt wird.

Im Beispiel:

$\frac{a}{b} = \frac{1}{2} = 0,5$ des Rohwertintervalls 11,5 bis 12,5, in dem die 15. Häufigkeit enthalten ist, werden zur Entsprechung der 15. Häufigkeit benötigt.

5. Addieren Sie zur unteren Grenze des in Schritt 2 bestimmten Rohwertintervalls den in Schritt 4 berechneten Intervallanteil. Dies ergibt den Median (Md).

Im Beispiel:

Md = 11,5 + 0,5
Md = 12,0.

Tab. 5:
Bestimmung des Medians für die Ergebnisse im Test Vitamine und Mineralstoffe.

In Tabelle 5 ist die Zahl der Rohwertträger mit $N = 30$ gerade. Tabelle 6 gibt ein fiktives Beispiel mit einer ungeraden Zahl von Rohwertträgern wieder ($N = 15$). Die Berechnung des Medians folgt den gleichen Schritten wie in Tabelle 5.

Rohwert (RW)	Häufigkeit (f)	cum _f	Bestimmung des Medians in Schritten
5	3	15	1. $\frac{1}{2} \cdot N = \frac{1}{2} \cdot 15 = 7,5$.
4	3	12	2. siehe Umrandung in Tabelle.
3,5	3	9	3. $7,5 - 5 = 2,5$ (a)
2,5	2	5	9 - 5 = 4 (b).
2	3	5	4. $\frac{(a)}{(b)} = \frac{2,5}{4} = 0,625$.
1	2	2	5. Median = $2,5 + 0,625 = 3,125$.
Summe	N = 15		

Tab. 6:

Fiktives Beispiel zur Berechnung des Medians, ungerades N, Schritte nach Tabelle 5, S. 22.

Übung 2:

Berechnen Sie für den Rohwertsatz in Tabelle 7 den Median. Folgen Sie bitte den Schritten nach Tabelle 6. Vor der Berechnung des Medians sind die cumulierten Häufigkeiten zu bestimmen. – Die Lösung ist auf der folgenden Seite angegeben.

Rohwert (RW)	Häufigkeit (f)	cum _f	Bestimmung des Medians in Schritten
50	1		1.
49	2		
48	2		2.
47	1		
46	0		3.
45	3		
44	3		4.
43	5		
42	1		5.
41	3		
40	2		
39	2		
38	1		
37	0		
36	3		
35	1		

Tab. 7:

Übungsbeispiel zur Berechnung des Medians, Lösung Seite 24.

Lösung der Übungsaufgabe 2:

Rohwert (RW)	Häufigkeit (f)	cum _f	Berechnung des Medians in Schritten
50	1	30	1. $\frac{1}{2} \cdot N = \frac{1}{2} \cdot 30 = 15.$
49	2	29	2. siehe Umrandung Tabelle.
48	2	27	3. $15 - 13 = 2$ (a)
47	1	25	$18 - 13 = 5$ (b).
46	0	24	4. $\frac{(a)}{(b)} = \frac{2}{5} = 0,40.$
45	3	24	5. $Md = 42,5 + 0,40$
44	3	21	$= 42,9.$
43,5	43	18	
42,5	42	13	
	41	3	
	40	2	
	39	2	
	38	1	
	37	0	
	36	3	
	35	1	
Summe	N = 30		

Tab. 8:

Lösung des Übungsbeispiels Tabelle 7.

Modalwert

Ein anderes Maß der zentralen Tendenz ist der Modalwert. Er ist das am einfachsten bestimmbare Maß zentraler Tendenz.

Definition 6:

Der *Modalwert* ist der Wert in einer Gruppe von Rohwerten, der am häufigsten vorkommt. Schreibweise: „Mo“.

Im Rohwertesatz

3, 3, 2, 2, 2, 4, 4, 4, 4, 5, 5, 6

ist 4 der Modalwert, da er häufiger als jeder andere Wert vorkommt. Die Bestimmung des Modalwertes ist unproblematisch, solange keine Sonderfälle auftreten. Solche Sonderfälle sind:

- (1) Alle Rohwerte treten gleich häufig auf:

11, 11, 14, 14, 16, 16

Für eine solche Gruppe von Werten gilt, daß sie keinen Modalwert hat.

- (2) Zwei aufeinanderfolgende Werte haben die gleiche Häufigkeit, und diese Häufigkeit ist größer als alle anderen Werte:

0, 0, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 5, 5, 6

Der Modalwert ist in diesem Falle das Mittel der aufeinanderfolgenden Werte: $Mo = (2+3)/2 = 2,5$.

- (3) Zwei nicht aufeinanderfolgende Werte haben die gleiche und die größte Häufigkeit:

5, 6, 6, 8, 8, 8, 10, 11, 12, 12, 12, 14, 14, 16

Es liegen zwei Modalwerte vor: $Mo = 8$ und $Mo = 12$. Eine derartige Verteilung wird *bimodal* genannt. – Treten in einer Verteilung deutlich zwei Häufungsspitzen auf, ist aber die Häufigkeit der Spitzen nicht genau gleich, spricht man im allgemeinen dennoch von bimodalen Verteilungen.

Für den Rohwertesatz in Tabelle 4, S. 18, beträgt der Modalwert 75. Der Rohwert 75 kommt mit 3 Besetzungen am häufigsten vor. – Der Modalwert liegt in dieser Verteilung noch verhältnismäßig nahe einem anderen, bereits besprochenen Maß der zentralen Tendenz: Der Median beträgt 80,5 (s. Seite 21).

Arithmetischer Mittelwert $\bar{x}$

Zwei Maße der zentralen Tendenz sind für einen Satz von N Rohwerten bereits bestimmt worden; der Median und der Modalwert. Es wird nun ein drittes Maß definiert, das arithmetische Mittel eines Satzes von Rohwerten. Dieser Mittelwert ist allgemein gemeint, wenn von dem „Durchschnitt“ gesprochen wird, wie z. B. von der „Durchschnittstemperatur“, von dem „durchschnittlichen Einkommen“ oder von dem „Durchschnittspreis“.

Definition 7:

Das *arithmetische Mittel*, kurz Mittel genannt, wird aus der Summe der Rohwerte dividiert durch die Anzahl der Rohwertträger gebildet.

Schreibweise: $\bar{x}$

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_N}{N} = \frac{\sum_{i=1}^N x_i}{N}$$

Lies für das Summenzeichen $\sum_{i=1}^N x_i$ (Griechisch Sigma):

"Summe über alle x_i von $i = 1$ bis N " oder kurz
 "Summe über x ".

Kurzschreibweise:

$$\bar{x} = \frac{\sum x}{N}$$

Sie werden wohl kaum noch einen arithmetischen Mittelwert für einen Satz von Rohwerten handschriftlich ausrechnen. In der Regel verfügen heute bessere, aber noch preiswerte elektronische Taschenrechner, zumindest wissenschaftlich-technische, über festprogrammierte statistische Funktionen wie Mittelwertberechnung $\bar{x}$ oder Berechnung der Standardabweichung s . Bei einem solchen Rechner brauchen nur die Rohwerte „eingetippt“ zu werden; durch entsprechenden Tastendruck erhalten Sie dann den Mittelwert oder die Standardabweichung (zur Standardabweichung siehe unten). Verfügen Sie über einen Rechner, der diese Funktionen nicht hat, empfiehlt es sich bei der Berechnung des Mittelwertes den Speicher des Taschenrechners zu benutzen. Dies ist besonders dann vorteilhaft, wenn Rohwerte mehrfach auftreten. – Über einen Speicher verfügt heute bereits der billigste Taschenrechner.

In Tabelle 9 ist der Mittelwert für die Ergebnisse im Test Vitamine und Mineralstoffe berechnet. Sie finden die Häufigkeitsverteilung der Rohwerte dieses Testes in Tabelle 3, S. 17. Die Berechnung wurde durchgeführt mit einem elektronischen Tischrechner der Firma Walther, Modell ETR 25 S. Dieser Tischrechner ist nach heutigen Ansprüchen primitiv, verfügt er doch neben einem Speicher praktisch nur über die vier Grundrechenarten. Der Vorteil dieses Rechners besteht darin, daß er den Rechengang ausdrückt.

An dem Ausdruck läßt sich der Rechengang verdeutlichen, der mit einem einfachen Taschenrechner mit Speicher vorgenommen werden kann. Aus Tabelle 3, S. 17, ist zu ersehen, daß der Rohwert 18 nur einmal vorkommt. Die Zahl 18 ist direkt in den Speicher eingetippt worden. Das Zeichen Σ neben der Zahl 18 zeigt an, daß der Wert 18 sich im Speicher befindet. Rohwert 17 kommt gleichfalls nur einmal vor und wird direkt in den Speicher eingetippt. Der Rohwert 16 ist zweimal vertreten. Es wird $2 \times 16 = 32$ gerechnet und danach die Speichertaste gedrückt. 32 erscheint im Speicher. Jede Zahl, die neu in den Speicher hinzukommt, wird

automatisch zum bereits vorhandenen Speicherinhalt aufaddiert. Sind so alle Rohwerte in den Speicher aufgenommen, läßt man den gesamten Speicherinhalt ausdrucken. 358 ist hier der Speicherinhalt Σ , wie aus der viertletzten Zeile in Tabelle 9 zu erkennen ist. Mit anderen Worten, die Summe der Rohwerte, die im Test Vitamine und Mineralstoffe erreicht wurde, ist 358. Für den arithmetischen Mittelwert ergibt sich, wenn diese Summe durch $N=30$ geteilt $\div$ wird, 11,93.

Bei elektronischen Tischrechnern, die in der Regel über einen Drucker verfügen, kann anhand des Ausdrucks nach der Rechnung geprüft werden, ob alle Werte richtig eingetippt worden sind. Bei elektronischen Taschenrechnern, die nur über eine Leuchtziffernanzeige verfügen, muß jeder Rechengang zur Kontrolle zweimal durchgeführt werden.

1 8 . 0 0	Σ	2 . 0 0	$\times$
1 7 . 0 0	Σ	1 2 . 0 0	$=$
2 . 0 0	$\times$	2 4 . 0 0	Σ
1 6 . 0 0	$=$	5 . 0 0	$\times$
3 2 . 0 0	$*$	1 1 . 0 0	$=$
3 2 . 0 0	Σ	5 5 . 0 0	Σ
2 . 0 0	$\times$	3 . 0 0	$\times$
1 5 . 0 0	$=$	1 0 . 0 0	$=$
3 0 . 0 0	$*$	3 0 . 0 0	Σ
3 0 . 0 0	Σ	4 . 0 0	$\times$
3 . 0 0	$\times$	8 . 0 0	$=$
1 4 . 0 0	$=$	3 2 . 0 0	Σ
4 2 . 0 0	$*$	7 . 0 0	Σ
4 2 . 0 0	Σ	6 . 0 0	Σ
5 . 0 0	$\times$	3 5 8 . 0 0	Σ
1 3 . 0 0	$=$	3 5 8 . 0 0	$\div$
6 5 . 0 0	$*$	3 0 . 0 0	$=$
6 5 . 0 0	Σ	1 1 . 9 3	$*$

Tab. 9:

Berechnung des Mittelwertes für den Rohwertsatz Tabelle 3, S. 17. Erklärung im Text.

Die Größe jedes Rohwertes in einer Verteilung von Rohwerten beeinflußt – im Gegensatz zum Median – den arithmetischen Mittelwert. Besonders sog. „Ausreißerwerte“, die an den Verteilungsenden liegen und mit dem eigentlichen Maß der zentralen Tendenz in keinem Zusammenhang mehr stehen, können den Mittelwert verfälschen. Dies sei an einem Zahlenbeispiel verdeutlicht:

		Mittel	Median	Modal
Rohwertsatz 1	1,3,4,5,5,6,7	$\frac{31}{7} = 4,43$	5	5
Rohwertsatz 2	1,3,4,5,5,6,21	$\frac{45}{7} = 6,43$	5	5

Im Rohwertsatz 2 ist der höchste Wert des Satzes 1 verdreifacht worden. Das Mittel verschiebt sich von 4,43 auf 6,43, während der Median und in diesem Falle auch der Modalwert gleich bleiben.

2.2 Symmetrische und schiefe Verteilungen

Für die Rohwerte des Lesereifetestes in Tabelle 4, S. 18, beträgt der Mittelwert $\bar{x} = 80,4$. Sie werden bemerken, daß für diese Verteilung der Median mit $M_d = 80,5$ (vgl. S. 21) nahezu mit dem Mittelwert identisch ist. Dies wird der Fall sein für jede annäherungsweise symmetrische Verteilung. Wird in einer Verteilung die linke Hälfte vom Mittelwert auf die rechte Hälfte geklappt und ergibt sich dabei eine exakte Deckung, liegt eine symmetrische Verteilung vor. Die Verteilungen vieler Persönlichkeitsmerkmale, wie z.B. Körpergröße, Gewicht, Brustumfang, Intelligenzquotient und Reaktionszeit tendieren dazu, symmetrisch und „Glocken-geformt“ zu sein. Verteilungen, die symmetrisch sind und einen Häufungsgipfel haben (unimodal), werden Normalverteilungen genannt. In einer vollkommenen Normalverteilung sind Mittelwert, Median und Modalwert gleich.

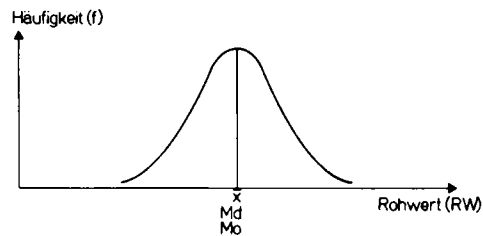


Abb. 1:
Symmetrische, unimodale Verteilung (Normalverteilung). Mittelwert, Median und Modalwert sind gleich.

Bei einem – aus welchen Gründen auch immer – für die Schüler leichtem Leistungstest wird es mehr hohe als niedrige Rohwerte geben. In diesem Falle erhalten Sie eine schiefe Verteilung. Der Häufungsgipfel ist nach rechts verschoben. Eine rechtsgipflige Verteilung wird auch *rechtsschief* oder negativ schief genannt.

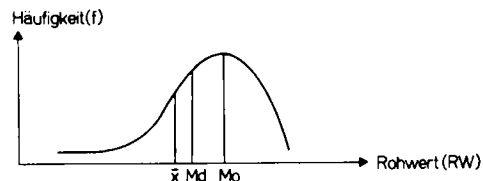


Abb. 2:
Rechtsschiefe Verteilung (leichter Test).

Bei einer rechtsschiefen Verteilung wird der Mittelwert von den niedrigen „Ausreißerwerten“ stark beeinflusst. Demzufolge ist der Mittelwert kleiner als der Median.

Entsprechend werden bei einem Test, der für die Schüler schwer zu bearbeiten ist, wenige hohe und viele niedrige Rohwerte vertreten sein. Auch in diesem Falle liegt eine schiefe, eine *linksschiefe* Verteilung vor. Sie wird auch positiv schief oder linksgipflig genannt. Hier ist der Mittelwert größer als der Median.

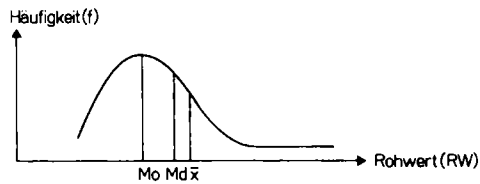


Abb. 3:
Linksschiefe Verteilung (schwerer Test).

Ohne eine Rohwertverteilung aufzuzeichnen, läßt sich aus den zwei Maßen der zentralen Tendenz Md und $\bar{x}$ bereits eine Aussage über die Verteilung machen. Ist $\bar{x}$ wesentlich kleiner als Md:

$\bar{x} \ll \text{Md}$, liegt eine rechtsschiefe Verteilung vor,
und ist $\bar{x}$ wesentlich größer als Md:

$\bar{x} \gg \text{Md}$, liegt eine linksschiefe Verteilung vor.

2.3 Die Auswahl eines Maßes zentraler Tendenz

Der arithmetische Mittelwert ist das zuverlässigste, d.h. stabilste Maß zentraler Tendenz. Zu seiner Berechnung wird jeder Rohwert in der Verteilung berücksichtigt.

Der Mittelwert sollte allerdings nicht benutzt werden, wenn der Umfang der Verteilung extrem klein ist. Bei einem N kleiner als 10 beeinflussen die „Ausreißerwerte“ zu hohem Anteil die Größe des Mittelwertes. Hier ist der Median besser geeignet.

Ferner wird der Mittelwert durch die Schiefe einer Verteilung stark beeinflusst, sei die Verteilung links- oder rechtsschief. In diesem Falle ist der Median gleichfalls ein besseres Maß zentraler Tendenz.

Mit anderen Worten, befinden sich in einer Gruppe von Rohwerten einige sehr extreme Werte, wird die zentrale Tendenz sinnvollerweise mit dem Median bestimmt. Jeder Rohwert einer Gruppe beeinflusst das arithmetische Mittel. Wenige Extremwerte können das Gruppenmittel weit aus dem Bereich herausdrücken, den man eigentlich als zentral für

die Rohwerte ansieht. Hierzu ein Beispiel: 90 Personen verfügen über Ersparnisse von 15 000 bis 16 700 DM mit einem Durchschnitt von 15 800 DM. Fügen Sie nun 10 Personen mit je einem Sparvermögen von 60 000 DM hinzu. Das durchschnittliche Sparvermögen dieser 100 Personen wird mit $\bar{x} = 20\,220$ DM irreführend hoch liegen, charakterisiert es doch nicht das Sparvermögen der überragenden Mehrheit der 90 Personen, die ein Sparvermögen von 15 000 bis 16 700 DM haben. Der Mittelwert wird weder der größeren noch der kleineren Gruppe gerecht. Deshalb zieht man hier den Median als Maß zentraler Tendenz vor. Der Median wird nicht beeinflusst durch die Größe der „großen“ und „kleinen“ Werte ober- und unterhalb von ihm. – Im obigen Beispiel wird der Median der kleineren Gruppe nicht gerecht. Es dürfte wohl sinnvoller sein, die kleinere Gruppe aus der Verteilung auszuschließen, für die größere Gruppe den Mittelwert anzugeben und das Sparvermögen der kleineren Gruppe getrennt davon mitzuteilen.

Manchmal kann es irreführend sein, für eine Verteilung ein zentrales Tendenzmaß anzugeben. Nehmen Sie an, von 5 Personen in einem Aufzug haben zwei 50 Pfennige, einer 50 DM, einer 200 DM und einer 5000 DM in der Tasche. Der Modalwert beträgt 50 Pfennige. Der Median liegt bei 50 DM und der Mittelwert wäre 1050,20 DM. Bei diesen Personen ist kein Maß zentral, da die Werte der Verteilung einfach nicht zentral tendieren.

Der Modalwert ist geeignet, mehrgipflige Verteilungen zu kennzeichnen. Auch bei mehrgipfligen Verteilungen ist es sinnlos, ein Zentralmaß anzugeben. Abbildung 4 verdeutlicht dies.

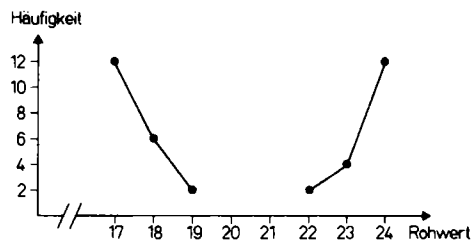


Abb. 4:

Bimodale Verteilung, $Mo_1 = 17$, $Mo_2 = 24$, $\bar{x} = 20,37$, $Md = 19$. Angabe eines Maßes zentraler Tendenz wie $\bar{x}$ und Md sinnlos.

Nach diesen Betrachtungen nun zurück zu der Rohwertverteilung im Beispielttest Vitamine und Mineralstoffe (Tabelle 3, S. 17). Für diese Verteilung ist $\bar{x} = 11,9$ und $Md = 12$. Aus der Häufigkeitsverteilung der Rohwerte dieses Testes können Sie ablesen, daß zwei Modalwerte auftre-

ten: $Mo_1 = 11$ und $Mo_2 = 13$. $\bar{x}$ und Mo sind nahezu gleich. Daraus folgt, daß die Rohwertverteilung nicht schief ist. Abbildung 5 gibt die Verteilung graphisch wieder.

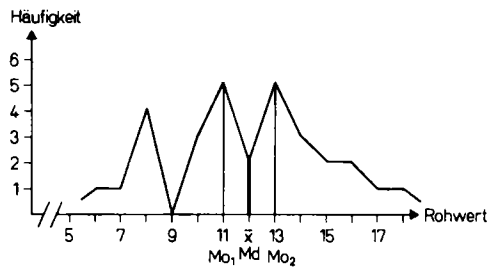


Abb. 5:
Häufigkeitsverteilung im Testbeispiel Vitamine und Mineralstoffe, Daten aus Tabelle 3, S. 17.

Von einer Normalverteilung in Abbildung 5 kann nur bedingt gesprochen werden. Mittelwert und Median sind zwar gleich, es liegen jedoch zwei Modalwerte vor. Die Verteilung ist bimodal. Allerdings liegen die zwei Häufungsspitzen dicht beieinander, so daß nicht von einer deutlichen bimodalen Verteilung gesprochen werden kann. Sie haben hier eine Häufigkeitsverteilung aus der praktischen Testentwicklung vor sich. Es sei daran erinnert, daß Abbildung 5 die Ergebnisse aus der nicht revidierten, ersten Fassung des Testes Vitamine und Mineralstoffe wiedergibt. Die Revision des Testes (siehe unten Aufgabenanalyse) wird dazu führen, daß bei einer wiederholten Testdurchführung in einer ähnlichen Klasse die zwei Häufungsspitzen verschwinden und eine mehr annähernd normale Verteilung entsteht.

3. Maße der Variabilität

Jedes Maß der zentralen Tendenz führt zu einem Wert, der auf bestimmte Weise stellvertretend für alle Werte eines Rohwertsatzes steht. Mit einem Zentralmaß ist jedoch nichts über den Grad der individuellen Unterschiede ausgedrückt, die zwischen den getesteten Personen bestehen können. Es werden statistische Kennwerte gebraucht, mit denen, wie man auch sagt, die Variabilität der Rohwerte in einer Gruppe von Werten beschrieben werden kann.

Es genügt also nicht, den Mittelwert oder den Median einer Gruppe von Testwerten zu kennen. Es muß auch etwas über die *Streuung* der Testwerte ausgesagt werden.

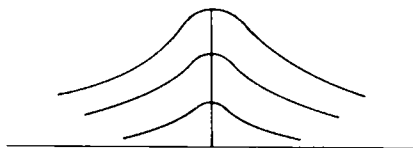


Abb. 6:
Verteilungen mit gleichem Mittelwert $\bar{x}$, Median M_d und Modalwert M_o , aber ungleicher Variabilität.

Die Maße der zentralen Tendenz ($\bar{x}$, M_d , M_o) sind bei allen drei Verteilungen in Abbildung 6 gleich. Im Grad der Variabilität oder Streuung unterscheiden sie sich jedoch erheblich. – Vorgestellt werden im folgenden drei Variabilitätsmaße: die Variationsweite, die Quartilabweichung und die Standardabweichung.

3.1 Variationsweite

Definition 8:

Die *Variationsweite* ist die Differenz zwischen dem größten und dem kleinsten Wert in einer Rohwertverteilung.

Schreibweise: V

$$V = R_{\max} - R_{\min}$$

mit

R Rohwert.

Sie können statt Variationsweite den englischen Ausdruck „range“ verwenden. Dieser Ausdruck ist im „Psychologen-Deutsch“ gebräuchlich.

Im Testbeispiel Vitamine und Mineralstoffe (Tabelle 3, S. 17) beträgt die Variationsweite $V = 18 - 6 = 12$. Die Spannweite der Werte, auf die sich die Leistung der Berufsfachschüler in diesem Test verteilt, beträgt 12 Rohwerte. Die höchstmögliche Variationsweite von 20 Rohwerten ist damit recht gut ausgeschöpft.

Die Variationsweite hat als Maß der Variabilität einen Nachteil. Die Variationsweite hängt allein von den Extremwerten in einer Verteilung ab. Die beiden äußersten Werte einer Verteilung können den Wert der Variationsweite erheblich beeinflussen. Somit ist die Variationsweite nur ein grobes Maß der Variabilität. Mit ihr wird wenig (oder u. U. gar nichts) über die tatsächlichen Verhältnisse in der Verteilung ausgesagt.

3.2 Quartilabweichung

Bevor auf die Quartilabweichung eingegangen werden kann, ist zuvor zu klären, was unter einem Quartil zu verstehen ist.

Definition 9:

Der Rohwert in einer Verteilung, unterhalb dessen 25 % der Werte liegen, ist das 1. Quartil (Q_1), der Rohwert unterhalb dessen 50 % der Werte liegen das 2. Quartil (Q_2), der Rohwert unterhalb dessen 75 % der Werte liegen das 3. Quartil (Q_3). – Ferner gilt:

$$P_{25} = Q_1, P_{50} = Q_2 \text{ u. } P_{75} = Q_3.$$

mit
P Perzentil.

Die Quartile bilden die drei Rohwerte, die eine Verteilung von Rohwerten in vier Teile zerlegen, wobei jeder Teil ein Viertel der Fälle enthält.

Die Entfernung zwischen dem ersten und dritten Quartil ($Q_3 - Q_1$) in einer Gruppe von Werten wird Interquartilweite genannt. Sie werden in Handanweisungen standardisierter Tests auch die Bezeichnung Interquartilsrange finden.

Die Quartilabweichung ist die Hälfte der Distanz $Q_3 - Q_1$.

Definition 10:

Die Quartilabweichung ist die Hälfte der Distanz zwischen dem 3. und 1. Quartil.

Schreibweise: Q

$$Q = \frac{Q_3 - Q_1}{2} = \frac{P_{75} - P_{25}}{2}$$

Anstelle der Bezeichnung Quartilabweichung wird auch die gelehrtete Bezeichnung Semi-Interquartilsrange verwendet.

Für den Test Vitamine und Mineralstoffe betragen die Perzentile P_{75} und P_{25} 14 bzw. 10 Rohwerte (siehe S.45 u. 44). Daraus läßt sich die Quartilabweichung für die Rohwerte angeben, die in diesem Test erreicht wurden:

$$q = \frac{Q_3 - Q_1}{2} = \frac{P_{75} - P_{25}}{2} = \frac{14 - 10}{2} = \frac{4}{2} = 2.$$

Die Interquartilweite $Q_3 - Q_1$ gibt die Variationsweite der mittleren 50% der Rohwerte wieder. Für die Rohwerte im Test Vitamine und Mineralstoffe beträgt dieser Wert (Zähler des obigen Ausdrucks) 4. Die mittleren 50% der Werte verteilen sich auf 4 Rohwerte⁴.

Quartilabweichung und Interquartilweite sind deutlich voneinander zu unterscheiden. Es gibt eine Quartilabweichung, auch Semi-Interquartils-

range genannt $\left(Q = \frac{Q_3 - Q_1}{2} \right)$, und eine Interquartilweite, auch Interquartilsrange genannt $(Q_3 - Q_1)$.

Überprüfen Sie nun zum Abschluß die Richtigkeit der zwei folgenden Aussagen:

- a) $Md = Q_2$
- b) In einer symmetrischen Verteilung muß gelten: $Q_1 + Q = Q_2$.

3.3 Standardabweichung

Das dritte und bekannteste Maß der Variabilität ist die Standardabweichung. Ähnlich wie beim Median und Modalwert geht auch bei der Bestimmung der Variationsweite und der Quartilabweichung nicht jeder Rohwert der Verteilung in diese Kennwerte ein. Die Standardabweichung ist ein Variabilitätsmaß, das von jedem Rohwert der Verteilung beeinflusst wird.

Definition 11:

Die *Standardabweichung* in einer Gruppe von Rohwerten ist die Wurzel aus dem Durchschnitt der quadrierten Abweichungen der Rohwerte vom Mittelwert.

Schreibweise: s

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{N}}$$

mit

- x Einzelrohwert,
- $\bar{x}$ Mittelwert des Rohwertsatzes,
- N Umfang des Rohwertsatzes.

Exkurs

An dieser Stelle ist ein statistischer Exkurs notwendig. Oftmals werden Sie Formeln zur Berechnung der Standardabweichung finden, bei denen die Summe der Abweichungsquadrate durch $n-1$ statt durch N geteilt wird.

Wenn eine Klasse von Schülern alle Personen umfaßt, über die Schlußfolgerungen gezogen werden sollen, dann stellt im statistischen Sinne die Schulklasse die Population dar. Arbeiten Sie mit Stichproben, soll die Populationsstandardabweichung genauer durch die Standardabweichung s Ihrer Stichprobe geschätzt werden, wenn die Summe der Abweichungsquadrate durch $n-1$ statt durch N dividiert wird.

Führen Sie hingegen Messungen an einer Population durch, und die Schulklasse stellt hier eine Population dar, ist die Summe der Abweichungsquadrate durch N zu teilen. In der Teststatistik, wie sie hier benötigt wird, soll aber kein Unterschied gemacht werden zwischen der Standardabweichung einer Stichprobe s und der Standardabweichung einer Population σ , aus der die Stichprobe stammt (σ ist der griechische Buchstabe „sigma“).

Die Werte der Standardabweichung werden sich ohnehin kaum unterscheiden, wenn bei ihrer Berechnung N oder $n-1$ verwendet wird, es sei denn N ist ausgesprochen klein ($N = 10$ oder kleiner). – Nach diesen statistischen Feinheiten zurück zur Berechnung der Standardabweichung nach Definition 11.

Für die Ergebnisse im Lesereifetest, Tabelle 4, S. 18, betrug der Mittelwert $\bar{x} = 80,4$. Nach der Formel in Definition 11 läßt sich die Standardabweichung wie folgt berechnen (vgl. Tabelle 4, S. 18):

$$\begin{aligned}s &= \sqrt{\frac{(112 - 80,4)^2 + (109 - 80,4)^2 + \dots + (44 - 80,4)^2}{36}} \\&= \sqrt{\frac{10.688,56}{36}} = \sqrt{296,9044} = \underline{17,2}\end{aligned}$$

Einfacher läßt sich die Standardabweichung mittels folgender „Maschinenformel“ berechnen. Es brauchen nicht mehr die Abweichungsquadrate gebildet zu werden:

$$s = \sqrt{\frac{\sum_N x^2 - (\sum_N x)^2}{N}}$$

Diese Formel wird benutzt, wenn ein einfacher elektronischer Taschenrechner oder Tischrechner zur Verfügung steht, der über einen Speicher verfügt. Für den Beispielttest Vitamine und Mineralstoffe ist in Tabelle 10

mit dieser Maschinenformel die Standardabweichung berechnet worden. Gerechnet wurde wieder auf dem elektronischen Tischrechner (Walther, ETR 25), der den Rechengang ausdruckt (vgl. S. 17).

18.00 x	256.00 Σ
18.00 =	7.00 x
324.00 *	7.00 =
324.00 Σ	49.00 *
17.00 x	49.00 Σ
17.00 =	6.00 x
289.00 *	6.00 =
289.00 Σ	36.00 *
16.00 x	36.00 Σ
16.00 x	4542.00 Σ
2.00 =	
512.00 *	
512.00 Σ	
15.00 x	
15.00 x	
2.00 =	
450.00 *	
450.00 Σ	
14.00 x	
14.00 x	
3.00 =	
588.00 *	
588.00 Σ	
13.00 x	
13.00 x	
5.00 =	
845.00 *	
845.00 Σ	
12.00 x	
12.00 x	
2.00 =	
288.00 *	
288.00 Σ	
11.00 x	
11.00 x	
5.00 =	
605.00 *	
605.00 Σ	
10.00 x	
10.00 x	
3.00 =	
300.00 *	
300.00 Σ	
8.00 x	
8.00 x	
4.00 =	
256.00 *	

↓

Σx^2

11.00 x

11.00 x

5.00 =

605.00 *

605.00 Σ

11² · 5 = 605

Der Rohwert 11 kommt 5 mal vor. Sein Quadrat wird mit 5 multipliziert. Das Ergebnis 605* wird in den Speicher aufgenommen: Σ

18.00	Σ	358.00	×
17.00	Σ	358.00	=
2.00	×	$(\Sigma x)^2 = 128164.00$	*
16.00	=		■
32.00	*		
32.00	Σ	30.00	×
2.00	×	4542.00	=
15.00	=	$N \cdot \Sigma x^2 = 136260.00$	*
30.00	*	136260.00	-
30.00	Σ	128164.00	=
3.00	×	$N \cdot \Sigma x^2 - (\Sigma x)^2 = 8096.00$	*
14.00	=	$\frac{8096.00}{8} = 1012.00$	✓ =
42.00	*	$\sqrt{1012} = 31.81$	*
42.00	Σ	89.98	÷
5.00	×	30.00	=
13.00	=	$\frac{18}{6} = 3.00$	*
65.00	*		
65.00	Σ		
2.00	×		
12.00	=		
24.00	*		
24.00	Σ		
5.00	×		
11.00	=		
55.00	*		
55.00	Σ		
3.00	×		
10.00	=	$s = \frac{\sqrt{N \Sigma x^2 - (\Sigma x)^2}}{N}$	
30.00	*		
30.00	Σ		
4.00	×		
8.00	=		
32.00	*		
32.00	Σ		
7.00	Σ		
6.00	Σ		
$\Sigma x = 358.00$			

Tab. 10:

Berechnung der Standardabweichung für die Ergebnisse im Test Vitamine und Mineralstoffe (Tabelle 3, S. 17) nach der „Maschinenformel“ Seite 35.

Technisch-wissenschaftliche Rechner haben heute in der Regel die Berechnung der Standardabweichung als statistische Funktion fest einprogrammiert. Es brauchen lediglich die Rohwerte eingetippt zu werden und durch entsprechenden Tastendruck erhält man die Standardabweichung. In Tabelle 10a ist dies für das gleiche Datenmaterial wie in Tabelle 10 demonstriert. Verwandt wurde ein elektronischer Taschenrechner der Marke Texas Instruments TI – 51 – III.

Tasten	Anzeige/Bemerkungen	Tasten	Anzeige/Bemerkungen
2nd → CA	0 Löschen		
18 → Σ+	1. Erste Eingabe	12 → Σ+	15. Fünfzehnte Eingabe
17 → Σ+	2. Zweite Eingabe	12 → Σ+	
16 Σ+	3. Dritte Eingabe	11 → Σ+	
16 → Σ+		11 → Σ+	
15 → Σ+		11 → Σ+	
15 → Σ+		11 → Σ+	
14 → Σ+		11 → Σ+	
14 → Σ+		10 → Σ+	
14 → Σ+		10 → Σ+	
13 → Σ+		10 → Σ+	
13 → Σ+		8 → Σ+	
13 → Σ+		8 → Σ+	
13 → Σ+		8 → Σ+	
13 → Σ+		8 → Σ+	
		7 → Σ+	
		6 → Σ+	30. Dreißigste Eingabe
		2nd → S.Dev.	3,0505 (Standardabweichung)
		2nd → Mean	11,933 (Mittelwert).

* Der Rohwert 16 ist zweimal vertreten und wird zweimal eingetippt.

Tab. 10a:

Berechnung der Standardabweichung und des Mittelwertes für den Test Vitamine und Mineralstoffe (Tabelle 3, S. 17). Taschenrechner: Texas Instruments TI – 51 – III (eingeprägte Funktionen: Standardabweichung und Mittelwert).

Tastenerläuterung:

2nd : (second) Tastaturebene 2

CA : clear all

Σ+ : Sigma Taste (Statistische Summierung)

S.Dev. : Standard Deviation (Standardabweichung)

Mean : Mittelwert

Mit $s = 3,0$ ist die Standardabweichung für die Ergebnisse im Test Vitamine und Mineralstoffe erheblich kleiner als für die Ergebnisse im Lesereifetest. Dies ist naheliegend: Der Bereich, über den die Rohwerte

im Test Vitamine und Mineralstoffe streuen, ist wesentlich kleiner als im Lesereifetest. – Bei der Besprechung der Normalverteilung (siehe unten) werden Sie ein Verfahren kennenlernen, wie die ungefähre Größe der Standardabweichung geschätzt werden kann, sofern die Rohwertverteilung bekannt ist.

Die Standardabweichung läßt sich geometrisch veranschaulichen⁵: In Tabelle 11 sind freigewählte Rohwerte von fiktiven Personen a–f aufgeführt. Der Mittelwert dieses fiktiven Rohwertsatzes beträgt $\bar{x} = 16$. In Spalte 3 der Tabelle 11 sind die Abweichungen der Rohwerte vom Mittelwert und in Spalte 4 die Quadrate dieser Abweichungen berechnet.

Der Durchschnitt der Abweichungsquadrate $\frac{\sum (x - \bar{x})^2}{N}$ beträgt $\frac{180}{6} = 30$. Die Wurzel daraus ergibt die Standardabweichung $s = 5,47$.

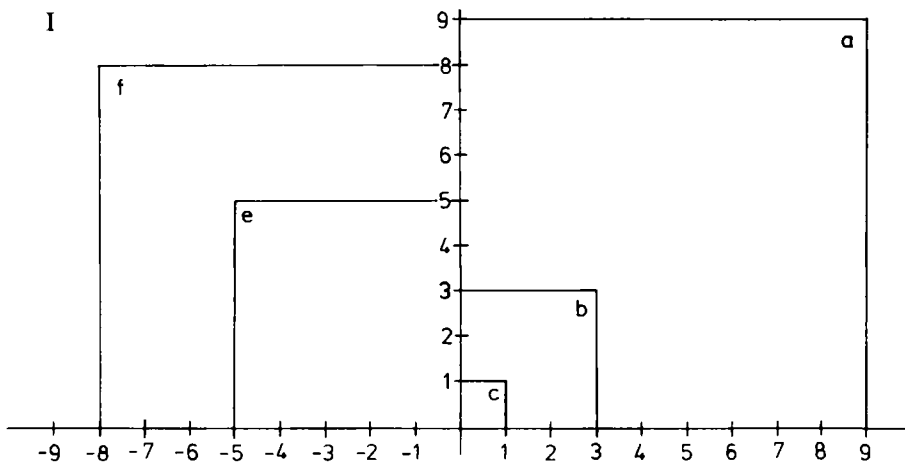
(1) Personen (N = 6)	(2) Rohwert x	(3) Abweichung x - $\bar{x}$	(4) Quadrierte Abweichung (x - $\bar{x}$) ²
a	25	+ 9	81
b	19	+ 3	9
c	17	+ 1	1
d	16	0	0
e	11	- 5	25
f	8	- 8	64
Summe	96	0	180
Mittelwert $\bar{x}$	16	-	30 = Varianz
Standardabweichung s	-	-	5,47

Tab. 11:
Daten zur geometrischen Interpretation der Standardabweichung in Abb. 7.

Der Rechengang in Tabelle 11 ist in Abbildung 7 geometrisch nachvollzogen. Die Teilabbildung I verdeutlicht Spalte 3 und 4 der Tabelle 11.

Person a weicht vom Mittelwert 9 Rohwerte ab. Das Abweichungsquadrat a deckt rechts von Null eine Fläche mit 81 Einheiten ab. Person f liegt 8 Rohwerte links vom Mittelwert. Das Abweichungsquadrat f liegt links von Null und umfaßt flächenmäßig 64 Einheiten. Die Summe der Quadrate a–f ergibt eine Fläche, die 180 Einheiten groß ist (Spalte 4, Tabelle 11). Diese Gesamtfläche, dividiert durch die Zahl der Quadrate (einschließlich des Nullquadrates), aus der sich die Gesamtfläche zusammensetzt, führt zu der durchschnittlichen Größe des Abweichungsquadrates:

$$\frac{180}{6} = 30. \text{ Teilabbildung II gibt dieses Quadrat wieder.}$$



II

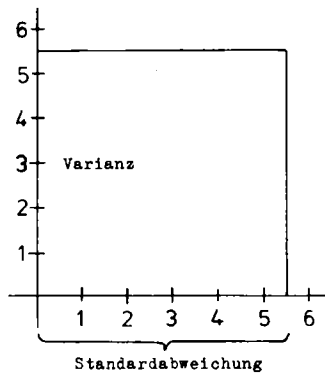


Abb. 7:

Geometrische Verdeutlichung: (i) der Abweichungen vom Mittelwert, (ii) der Quadrate dieser Abweichungen, (iii) des Mittelwertes der Abweichungsquadrate (Varianz) und (iv) der Standardabweichung (Daten aus Tab. 11).

Die Seitenlänge des Quadrates in Teilabbildung II ist die Standardabweichung $s = \sqrt{30} = 5,47$. Die Fläche dieses Quadrates, s^2 , wird mit Varianz bezeichnet. Es gilt:

$$s^2 = \text{Fläche des durchschnittlichen Abweichungsquadrates} = \text{Varianz.}$$

$$s = \sqrt{\text{Fläche des durchschnittlichen Abweichungsquadrates}} = \sqrt{\text{Varianz.}}$$

Definition 12:

Die *Varianz* ist das Quadrat der Standardabweichung.
Schreibweise: s^2 .

Die Standardabweichung bzw. Varianz wird wie das arithmetische Mittel von jedem Rohwert in einer Gruppe von Rohwerten beeinflusst. Die Variabilitätsverhältnisse einer Verteilung gibt demzufolge die Standardabweichung im allgemeinen exakter wieder als die Quartilabweichung; in die Bestimmung der Quartilabweichung geht nicht jeder Meßwert ein.

Die Standardabweichung wird jedoch durch die Tatsache, daß alle Meßwerte in ihre Bestimmung eingehen, auch durch die extremen Werte einer Verteilung beeinflusst. Was dies bedeuten kann, ersehen Sie aus Abbildung 7. Extreme Werte sind weit vom Mittelwert entfernt. Gehen sie wie die anderen Werte in das Quadrat bei der Berechnung der Standardabweichung ein, können sie überrepräsentiert sein. Denken Sie sich für Abbildung 7 eine Person, die 18 Rohwerte vom Mittelwert entfernt ist. Die entsprechende Quadratfläche mit 324 Einheiten würde die durchschnittliche Quadratfläche der Abweichungen, d.h. die Varianz, erheblich vergrößern. Die Quartilabweichung wird dagegen durch das Vorliegen extremer Werte weniger beeinflusst:

	Rohwertsätze	s	Q
I	15, 18, 19, 20, 22, 25, 26, 29	4,7	3,5
II	15, 18, 19, 20, 22, 25, 26, 29, 39	7,2	3,8

Zum zweiten Rohwertsatz ist ein Extremwert hinzugefügt worden. Durch das Hinzukommen des Extremwertes wird die Quartilabweichung weitaus weniger verändert als die Standardabweichung.

Die Standardabweichung ist weniger als Variabilitätsmaß geeignet, wenn die Rohwerte einige sehr extreme Werte aufweisen oder wenn die Verteilung ausgesprochen schief ist. Wenn der Umfang der Rohwertverteilung jedoch groß ist ($N > 25$) und nur einige wenige extreme Werte auftreten, dürfte dieser Nachteil der Standardabweichung kaum ins Gewicht fallen.

Hat sich bei der Bestimmung des Maßes zentraler Tendenz herausgestellt, daß für eine bestimmte Rohwertverteilung der Median geeigneter ist als der arithmetische Mittelwert, empfiehlt es sich, als passendes Maß der Variabilität die Quartilabweichung zu verwenden. Für den Median ist die Entsprechung die Quartilabweichung, für den arithmetischen Mittelwert ist es die Standardabweichung.

Die Standardabweichung gewinnt besondere Bedeutung bei der Inter-

pretation der Leistung, die eine Person in einem Test erreicht. Die Position einer Person in einer Verteilung wird oft in bezug auf die Standardabweichung ausgedrückt. In dem Test Vitamine und Mineralstoffe betrug der Mittelwert aufgerundet $\bar{x} = 12$ und die Standardabweichung $s = 3$. Ein Schüler, der einen Rohwert von $RW = 15$ erreichte, liegt eine Standardabweichung über dem Mittelwert. Für seinen Rohwert kann $\bar{x} + 1 \cdot s$ geschrieben werden. In gleicher Weise kann für den Rohwert eines Schülers, der 9 Rohwerte erreichte, $\bar{x} - 1 \cdot s$ geschrieben werden. Sein Rohwert liegt eine Standardabweichung unter dem Mittelwert. Auf diese Verwendung der Standardabweichung wird noch später eingegangen.

4. Perzentile und Prozentränge

Perzentile

Es sollen nun die Leistungen, die im Test Vitamine und Mineralstoffe erreicht wurden, in drei Gruppen aufgeteilt werden:

Gruppe I Schwache Leistung

Gruppe II Durchschnittliche Leistung

Gruppe III Hohe Leistung

Diese Gruppen sollen in folgender Weise festgelegt werden:

I Schwache Leistung: die unteren 25 % der Rohwertverteilung

II Durchschnittliche Leistung: die mittleren 50 % der Rohwertverteilung

III Hohe Leistung: die oberen 25 % der Rohwertverteilung

Zur Lösung dieses Problems sind zwei Punkte in der Rohwertskala zu suchen:

a der Rohwert, unterhalb dessen die Werte eines Viertels der Verteilung liegen und

b der Rohwert, unterhalb dessen die Werte von dreiviertel der Verteilung liegen.

Zwischen diesen beiden Punkten auf der Rohwertskala müssen die mittleren 50 % der Verteilung liegen. Diese gesuchten Punkte bezeichnet man als *Perzentile*. Gesucht sind hier das Perzentil 25 und das Perzentil 75. Kurz ausgedrückt schreibt man

Perzentil 25 als P_{25}

Perzentil 75 als P_{75}

Definition 13:

Das *Perzentil* P_p ist derjenige Punkt in einer Rohwertverteilung, unterhalb dessen p Prozent der in einer Primären Häufigkeitsverteilung angeordneten Rohwerte einer Gruppe liegen.

Schreibweise: P_p , z. B.: P_{30}

Nun ist der Median ein bestimmtes Perzentil. Er ist das 50. Perzentil in einer Gruppe von Rohwerten; es ist $Md = P_{50}$. Bei der Berechnung eines Perzentils wird ähnlich vorgegangen wie bei der Berechnung des Medians (siehe Tabelle 5, S. 22). Tabelle 12 zeigt das Vorgehen für die Berechnung des 25. Perzentils. Statt $(1/2) \cdot N$ wie bei der Medianberechnung in Schritt 1 müssen Sie nun N mit 0,25 oder mit $1/4$ multiplizieren.

Rohwert (RW)	Häufigkeit (f)	cum _f	Schritte zur Bestimmung des 25. Perzentils
18	1	30	1. $0,25 \cdot N = 1/4 \cdot 30 = 7,5$.
17	1	29	2. Suchen Sie in der Spalte cum _f das Intervall auf, das die 7,5. Häufigkeit enthält. (siehe Umrandung)
16	2	28	3. Die Differenz zwischen der 7,5. Häufigkeit und cum _f = 6 ist
15	2	26	$7,5 - 6 = 1,5$ (a)
14	3	24	Die Differenz zwischen cum _f = 9 und cum _f = 6 ist
13	5	21	$9 - 6 = 3$ (b)
12	2	16	4. $\frac{(a)}{(b)} = \frac{1,5}{3} = 0,5$ des Intervalls werden benötigt.
11	5	14	5. $P_{25} = 9,5 + 0,5 = 10,0$.
10,5	3	9	
9,5	0	6	
9	4	6	
8	1	2	
7	1	1	
6	1	1	
Summe	N = 30		

Tab. 12:

Bestimmung des 25. Perzentils für die Ergebnisse im Beispieltest Vitamine und Mineralstoffe.

Unter dem Rohwert RW = 10,0 liegen somit 25% der Verteilung. Die mittleren 50% der Verteilung liegen zwischen den Rohwerten RW = 10,0 und RW = 14,0 (zwischen P_{25} und P_{75}). Über RW = 14,0 liegen die oberen 25% der Verteilung. Die Berufsfachschüler mit schwacher Leistung im Test Vitamine und Mineralstoffe erreichen höchstens 10, die mit durchschnittlicher Leistung zwischen 10 und 14 und die mit hoher Leistung über 14 Rohwerte.

Übung 3:

Berechnen Sie bitte für den Rohwertesatz in Tabelle 12 das Perzentil P_{75} . Folgen Sie den 5 Schritten nach Tabelle 12. Die Lösung ist auf der nächsten Seite angegeben.

Schritt

- 1.
2. Benutzen Sie die Verteilung in Tabelle 12
- 3.
- 4.
5. $P_{75} =$

Lösung der Übungsaufgabe 3:

$$1. \quad 0,75 \cdot N = 0,75 \cdot 30 = 22,5$$

$$4. \quad \frac{(a)}{(b)} = \frac{1,5}{3} = 0,5$$

$$2. \quad \text{Intervall: } \begin{array}{c} 14,5 \\ 13,5 \end{array} \begin{array}{c} \text{RW} \\ \text{f} \end{array} \begin{array}{|c|c|} \hline 14 & 24 \\ \hline \end{array}$$

$$5. \quad P_{75} = 13,5 + 0,5 = 14,0$$

$$3. \quad \begin{array}{l} 22,5 - 21 = 1,5 \quad (a) \\ 24 - 21 = 3 \quad (b) \end{array}$$

Prozentränge

Bei der Berechnung eines Perzentils geht es darum, denjenigen Rohwert aufzusuchen, unterhalb dessen p Prozent der Rohwerte liegen. Dies geht auch umgekehrt: Ein bestimmter Rohwert ist gegeben und es soll bestimmt werden, wieviel Prozent der Rohwerte unterhalb dieses Wertes liegen. Dieser Wert wird als Prozentrang eines Rohwertes bezeichnet.

Definition 14:

Der *Prozentrang* eines Rohwertes x gibt den Prozentsatz der Rohwerte an, die gleich oder kleiner als der Rohwert x sind.

Schreibweise: PR

$$PR = \frac{\text{cum}_f}{N} \cdot 100 \quad [\%]$$

Prozentränge und Perzentile dürfen nicht verwechselt werden. Bei Perzentilen geht es darum, zu einem vorgegebenen Prozentwert den entsprechenden Rohwert in der Verteilung aufzusuchen. Beim Prozentrang wird ausgehend von einem Rohwert der entsprechende Prozentwert gesucht. Die Bestimmung des Perzentils führt zu einem Rohwert, die des Prozentranges zu einem Prozentwert.

In Tabelle 13 sind die Prozentränge für die Rohwerte bestimmt, die von den Berufsfachschülern im Test Vitamine und Mineralstoffe erzielt wurden.

Zur Berechnung der Prozentränge sind zuerst die cumulierten Häufigkeiten zu bestimmen (siehe cum_f S. 21). In jeder Zeile wird die aufsummierte Häufigkeit durch die Anzahl der Rohwertträger N dividiert und dieser Quotient mit 100 multipliziert, um Prozentwerte zu erhalten. Prozentränge veranschaulichen die Position eines Rohwertes und damit einer Person in einer Rohwertverteilung. Zu jedem Rohwert in Tabelle 13 läßt sich der entsprechende Prozentrang aufsuchen. Z. B. hat der Rohwert 11 den Prozentrang 46,66, d. h. aufgerundet den Prozentrang 47. Dies be-

Rohwert (RW)	Häufigkeit (f)	cum _f	Prozentrang (PR) (%)
18	1	30	100
17	1	29	96,66
16	2	28	93,33
15	2	26	86,66
14	3	24	80
13	5	21	70
12	2	16	53,33
11	5	14	46,66
10	3	9	30
9	0	6	20
8	4	6	20
7	1	2	6,66
6	1	1	3,33

Summe N = 30

Tab. 13:

Bestimmung der Prozentränge (PR) für die Rohwerte im Test Vitamine und Mineralstoffe.

deutet, daß 47 % der Schüler höchstens den Rohwert 11 erreichen. Oder anders ausgedrückt: Erreicht ein Schüler 11 Rohwerte, hat der den Prozentrang 47. Sein Rohwert ist gleich oder höher als 47% der Mitschüler. 53 % erreichen dagegen einen höheren Rohwert. Der Rohwert 14 hat bereits den Prozentrang 80. Nur 20 % der Schüler erzielen einen höheren Rohwert.

5. Standardnormalverteilung (z-Verteilung)

Normalverteilung

Die Normalverteilungskurve oder Normalkurve beschreibt eine spezielle, symmetrische, unimodale Verteilung. Normalverteilungen mit verschiedenem $\bar{x}$ und s zeigt Abbildung 8.

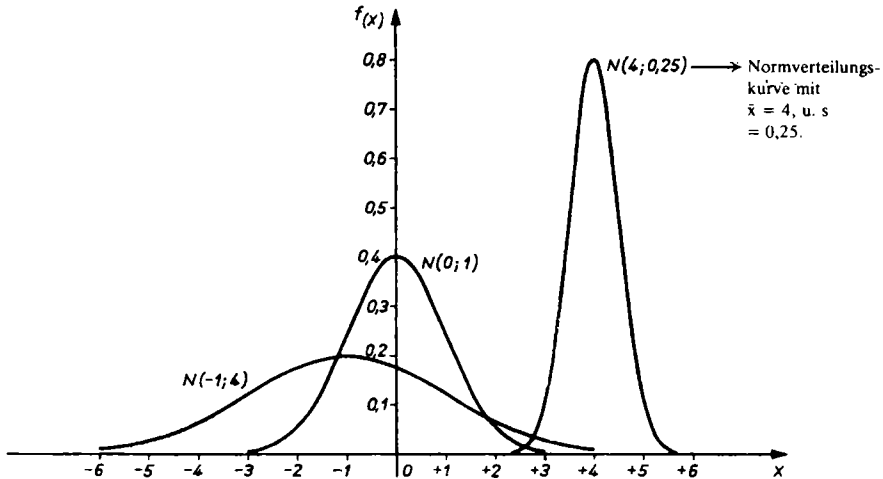


Abb. 8:

Normalverteilungen mit verschiedenen Mittelwerten $\bar{x}$ und Standardabweichungen s (G. Clauß, H. Ebner, Grundlagen der Statistik für Psychologen, Pädagogen und Soziologen, ver. u. erw. Aufl., Berlin 1974, S. 152).

Die wesentlichen Eigenschaften der Normalverteilung sind:

- Die Kurve ist symmetrisch um $\bar{x}$.
- Die Verteilung ist unimodal.
- Die Kurve hat Wendepunkte⁶ bei $\bar{x} \pm s$.
- Das Maximum liegt bei $\bar{x}$. Extreme Abweichungen vom Mittelwert haben eine außerordentlich geringe Auftretenswahrscheinlichkeit.

Die Normalverteilungskurve ist mathematisch über folgende Funktionsgleichung definiert:

$$f(x) = \frac{1}{\sqrt{2\pi} \cdot s} \cdot e^{-1/2 \left(\frac{x - \bar{x}}{s} \right)^2}$$

Das Symbol e steht für die Eulersche Zahl 2,718... Die Zahl „Pi“, $\pi = 3,141...$ ist noch aus den Anfangsgründen der Geometrie bekannt. s und $\bar{x}$ sind die Standardabweichung und der Mittelwert der in Frage stehenden Normalverteilung.

Die exakte Formel für die Normalverteilung bleibt im folgenden ohne Bedeutung. Von besonderer Bedeutung ist die Standardnormalverteilung oder z-Verteilung.

Standardnormalverteilung oder z-Verteilung

Die Standardnormalverteilung hat einen Mittelwert von $\bar{x} = 0$ und eine Standardabweichung von $s = 1$. In Abbildung 8 ist es die Normalverteilungskurve, die mit $N(0,1)$ gekennzeichnet ist. Es ist dies die Verteilung die gewöhnlich als Standardnormalverteilung oder z-Verteilung in Statistik-Tabellen erscheint. Abbildung 9 gibt die Standardnormalverteilung (z-Verteilung) in einem größeren Maßstab wieder⁷. Die Fläche unter der Kurve steht für eine Wahrscheinlichkeit von 1,00; die Wahrscheinlichkeit des Auftretens von z-Werten im Intervall von $-3,00$ bis $+3,00$ beträgt 1,00 oder 100%.

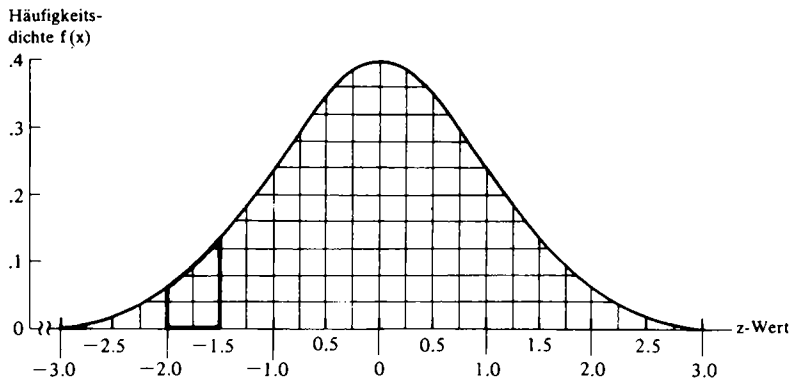
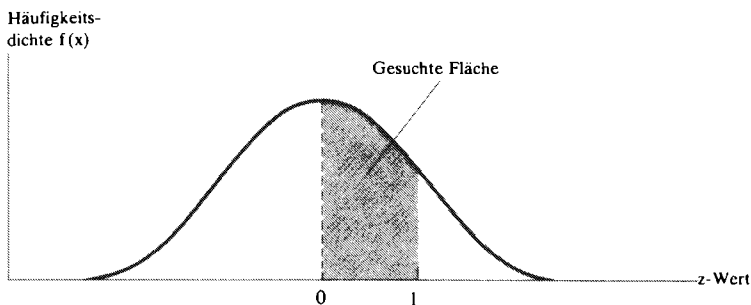


Abb. 9:

Standardnormalverteilung $N(0,1)$. Die gesamte Fläche unter der Kurve repräsentiert eine Wahrscheinlichkeit von 1,00. Die Gesamtfläche umfaßt 100 Quadrate. Jedes einzelne ganze Quadrat steht für eine Wahrscheinlichkeit von 0,01 (E. P. Lo-vejoy, Statistics for math haters, New York 1975, S. 141).

Die Kurve der Standardnormalverteilung in Abbildung 9 ist so gezeichnet, daß die gesamte Fläche, die sie abdeckt, 100 Quadrate umfaßt. Jedes ganze Quadrat steht für eine Wahrscheinlichkeit von $1/100$ oder 0,01. Bei Verwendung dieser Abbildung können Sie für jedes Intervall, das Sie interessiert, die Auftretenswahrscheinlichkeit aufsuchen. Es soll z. B. die Wahrscheinlichkeit des Auftretens von z-Werten bestimmt werden, die zwischen $z = 0$ und $z = +1$ liegen:

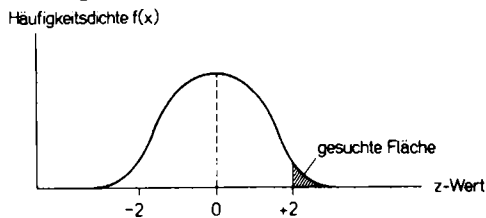


Sie können in Abbildung 9 auszählen, daß ca. 34 Quadrate zwischen $z = 0$ und $z = 1$ liegen. Da ein Quadrat für eine Wahrscheinlichkeit von 0,01 steht, muß die Auftretenswahrscheinlichkeit für einen z -Wert zwischen $z = 0$ und $z = 1$: $34 \cdot 0,01 = 0,34$ betragen. Das heißt, 34 % aller z -Werte liegen im Intervall von $z = 0$ bis $z = 1$. Da die Normalverteilung symmetrisch ist, beträgt die Auftretenswahrscheinlichkeit für einen Wert zwischen $z = -1$ und $z = 0$ gleichfalls 0,34. Im Intervall von $z = -1$ bis $z = +1$ beträgt die Wahrscheinlichkeit des Auftretens von Werten: $0,34 + 0,34 = 0,68$. Das heißt, 68 % aller Fälle oder Werte liegen im Intervall von $z = -1$ bis $z = +1$.

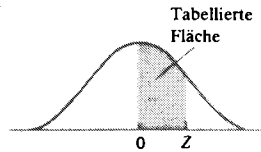
5.1 Tabelle

Die Methode des Quadratauszählens ist umständlich und ungenau. Man hat über eine exakte Ausrechnung nach der mathematischen Funktionsgleichung für die Standardnormalverteilung eine Tabelle angelegt, aus der die Auftretenswahrscheinlichkeit eines Wertes zwischen Null und einem gegebenen z -Wert aufgesucht werden kann. Auf Seite 50 ist eine solche Tabelle für die Standardnormalverteilung wiedergegeben (z -Tabelle). Unter $z = 1$ finden Sie den Wert 0,3413. Dies entspricht dem ungefähren Wert, der über das Auszählen der Quadrate in Abbildung 9 gefunden wurde.

Beispiel 1: Es soll die Auftretenswahrscheinlichkeit für einen z -Wert größer als $z = +2$ gefunden werden:



Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4773	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4983	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4989	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998
3.6	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.7	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.8	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.5000
3.9	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09

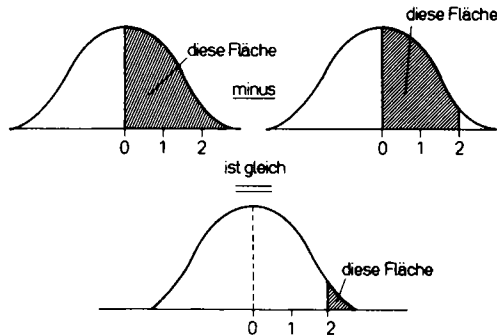


Beispiel: Die Fläche korrespondierend zu einem z-Wert von 1,96 ist 0,4750. Demzufolge ist die Wahrscheinlichkeit für einen z-Wert zwischen 0 und 1,96 0,4750. 47,5% aller Fälle oder Werte liegen im Intervall $z = 0$ bis $z = 1,96$.

Tab. 14:

Flächentabelle für die Standardnormalverteilung, z-Tabelle (E. P. Lovejoy, Statistics for math haters, New York 1975, S.143).

Es muß die Differenz zweier Flächen gebildet werden:



Nach Tabelle 14 gilt für $z = +3,99$ und $z = +2$:
 $0,5000 - 0,4773 = 0,0227$.

Die Wahrscheinlichkeit, einen Wert größer als $z = +2$ zu erhalten, beträgt 0,0227. Das heißt, 2,27 % aller Werte oder Fälle in einer Normalverteilung sind größer als $z = 2$. Die Normalverteilungskurve ist symmetrisch. Demzufolge ist auch die Wahrscheinlichkeit für einen Wert kleiner als $z = -2$ gleich 0,0227. Das heißt wiederum, daß 2,27 % aller Werte in einer Normalverteilung links von $z = -2$ liegen.

Übung 4:

Bestimmen Sie bitte die Wahrscheinlichkeit des Auftretens eines Wertes zwischen:

- a) $z = 0$ und $z = +1,5$
- b) $z = 0$ und $z = -1,5$
- c) $z = 1$ und $z = +2,0$
- d) $z = -1,43$ und $z = +0,78$

Die Lösungen sind auf Seite 52 angegeben.

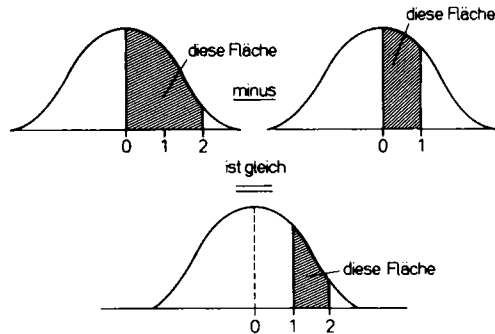
Tabelle 14 läßt sich auch umgekehrt benutzen. In welchen z -Wertebereich fallen 25 % aller Werte rechts vom Mittelwert $z = 0$? Wenn Sie die Auftretenswahrscheinlichkeit 0,25 in Tabelle 14 aufsuchen, finden Sie den zugehörigen z -Wert $z = +0,67$ (ohne Interpolation). Zwischen $z = 0$ und $z = +0,67$ liegen 25 % der Werte rechts vom Mittelwert $z = 0$. Entsprechend liegen zwischen $z = 0$ und $z = -0,67$ links vom Mittelwert 25 % aller Werte. Dies bedeutet, daß die mittleren 50 % der Werte oder Fälle in einer Normalverteilung zwischen $z = -0,67$ und $z = +0,67$ liegen.

Übung 5:

Ermitteln Sie bitte den z -Wertebereich, in dem die mittleren 68 % der Werte oder Fälle einer Normalverteilung liegen. Die Lösung ist auf Seite 53 angegeben.

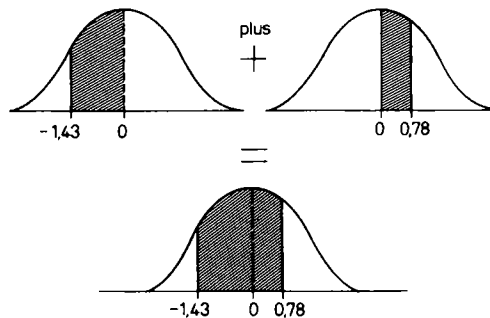
Lösung der Übungsaufgabe 4:

- In Tabelle 14 findet sich für den z-Wert $z = +1,5$ der Wert 0,4332. 43,32 % aller Werte oder Fälle liegen im Intervall $z = 0$ bis $z = 1,5$. Die Auftretenswahrscheinlichkeit beträgt 0,4332.
- Da die Normalverteilung symmetrisch ist, liegen im Intervall von $z = 0$ bis $z = -1,5$ gleich viele Werte oder Fälle wie im Intervall von $z = 0$ bis $z = 1,5$. Die Auftretenswahrscheinlichkeit beträgt wie bei Teilaufgabe a) 0,4332.
- Folgende Flächen müssen subtrahiert werden:



Die Fläche zwischen 0 und 2 (Tabelle 14, $z = 2$) beträgt 0,4773. Die Fläche zwischen 0 und 1 (Tabelle 14, $z = 1$) beträgt 0,3413. $0,4773 - 0,3413 = 0,1360$. Die Auftretenswahrscheinlichkeit für einen z-Wert zwischen $z = 1$ und $z = 2$ ist 0,1360. 13,60 % der Werte oder Fälle in einer Normalverteilung liegen zwischen $z = 1$ und $z = 2$.

- Da die z-Werte $-1,43$ und $0,78$ links und rechts des Mittelwertes liegen, müssen Flächen addiert werden:



$$\begin{aligned} \text{Fläche zwischen} \\ -1,43 \text{ und } 0,78 &= 0,4236 + 0,2823 \\ &= 0,7059 \end{aligned}$$

Die Auftretenswahrscheinlichkeit für z-Werte zwischen $-1,43$ und $+0,78$ beträgt $0,7059$. $70,59\%$ aller Werte oder Fälle in einer Normalverteilung liegen zwischen diesen z-Werten.

Lösung der Übungsaufgabe 5:

Gleich große Flächen liegen beiderseits des Mittelwertes: $\frac{68\%}{2} = 34\%$.

Aus Tabelle 14 ergibt sich für die Auftretenswahrscheinlichkeit $0,34$ ein z-Wert von $z = 1,0$ (ohne Interpolation). Die Normalverteilung ist symmetrisch. Demzufolge liegen im z-Bereich von $z = -1$ bis $z = +1$ 68% der Werte oder Fälle.

5.2 Standardwerte z

In Übung 5 wurde für verschiedene Intervalle der Standardnormalverteilung die Auftretenswahrscheinlichkeit von Werten oder Fällen bestimmt. Von besonderem Interesse ist die Auftretenswahrscheinlichkeit oder Häufigkeit von Fällen zwischen den ganzzahligen z-Werten, z. B. zwischen $z = -1$ und $z = +1$ oder zwischen $z = +1$ und $z = +2$. Bei der Standardnormalverteilung oder z-Verteilung $N(0,1)$ beträgt die Standardabweichung $s = 1$, so daß die ganzzahligen z-Werte $1, 2, 3$ und $-1, -2, -3$ die Normalverteilung in Einheiten der Standardabweichungen aufteilen. Für den z-Wert $z = +1$ läßt sich nämlich $\bar{x} + 1 \cdot s$ schreiben, da folgendes gilt: $z = \bar{x} + 1 \cdot s = 0 + 1 \cdot 1 = 1$. Abbildung 10 zeigt die Standardnormalverteilung eingeteilt in ganzzahlige z-Werte, d. h. eingeteilt in Einheiten der Standardabweichung. Es ist mit angegeben wieviel Prozent der Fälle zwischen den jeweiligen z-Werten bzw. Standardabweichungen liegen. Über Tabelle 14 (Kap. 5.1) lassen sich die angegebenen Prozentangaben nachprüfen.

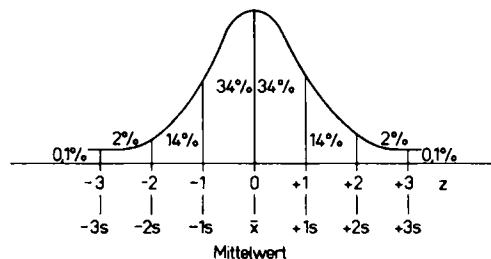


Abb. 10:
Standardnormalverteilung, eingeteilt in Einheiten der Standardabweichung oder ganzzahlige z-Werte.

In einer Standardnormalverteilung liegen 68% der Werte im Bereich von $-1 \cdot s$ bis $+1 \cdot s$. Dies sind ca. $2/3$ aller Fälle. $1/3$ aller Fälle liegen demnach links und rechts dieses Bereiches. Das bedeutet, daß $1/6$ (die Hälfte von $1/3$) aller Werte über $+1 \cdot s$ und $1/6$ unter $-1 \cdot s$ liegen (Abbildung 11).

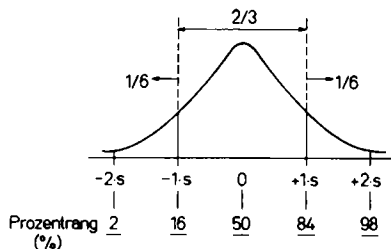


Abb. 11:
Standardnormalverteilung und Prozentränge.

In der Standardnormalverteilung fallen wie in jeder Normalverteilung Mittelwert, Median und Modalwert zusammen. Der Median ist das 50. Perzentil und liegt beim Mittelwert $z = 0$. 50% aller Werte liegen unter $z = 0$. Aus Abbildung 10, Seite 53, sehen Sie, wieviel Prozent der Werte zwischen $z = 0$ und $z = +1$, bzw. zwischen 0 und $+1 \cdot s$ liegen. Daraus läßt sich der Prozentrang für den z -Wert $z = +1$ bestimmen. Von links ausgehend bis zum Median liegen 50% der Werte. Bis $+1 \cdot s$ oder bis $z = +1$ liegen dann $50\% + 34\% = 84\%$. Der Prozentrang eines Wertes, der eine Standardabweichung über dem Mittelwert $z = 0$ liegt, beträgt somit 84. Aus Abbildung 11 sind die Prozentränge für die übrigen markanten Werte der Standardnormalverteilung zu sehen. Z.B. erreichen höchstens 16% ($50\% - 34\% = 16\%$) der Probanden in einer Normalverteilung Werte, die links der ersten negativen Standardabweichung liegen (links von $-1 \cdot s$ oder $z = -1$). – Die hier gemachten Prozentangaben gelten nicht nur für die Standardnormalverteilung sondern auch für jede Normalverteilung, die in Einheiten ihrer Standardabweichung eingeteilt wird.

Die Rohwerte eines Testes lassen sich nun gut interpretieren, wenn sie in Einheiten der Standardabweichung ausgedrückt werden: Das heißt, die Rohwertabweichung vom Mittelwert wird durch die Standardabweichung dividiert. An einem Beispiel sei dies verdeutlicht: Eine Person erreicht einen Testwert, der 12 Rohwerte unter dem Mittelwert liegt, wobei die Rohwerte normal verteilt sind. Wie tief liegt der Rohwert dieser Person wirklich? Ist die Standardabweichung klein ($s = 4$) mag es der kleinste Rohwert unter 100 getesteten Personen sein ($-3 \cdot s$) mit einem Prozentrang von weniger als 1. Keiner der getesteten Personen wird mehr als 12 Rohwerte unter dem Mittelwert liegen. Wenn die Standardabweichung größer wäre und 12 Rohwerte betragen würde, würde diese Leistung ge-

nau eine Standardabweichung links vom Mittelwert liegen ($-1 \cdot s$). Aus Abbildung 11 ist zu ersehen, daß der Wert $-1 \cdot s$ den Prozentrang 16 hat. 16% der getesteten Personen würden nun mehr als 12 Rohwerte unter dem Mittelwert liegen. – Man bestimmt einen „Standardrohwert“, kurz „Standardwert“, der anzeigt, wieviele Standardabweichungen der Rohwert einer Person über oder unter dem Mittelwert liegt. Der Grund-Standardwert ist der z-Wert.

Definition 15:

Der *Standardwert* z einer Person ist der Quotient aus der Rohwertabweichung vom Mittelwert und der Standardabweichung.

$$z = \frac{\text{Rohwert} - \text{Mittelwert}}{\text{Standardabweichung}} = \frac{x - \bar{x}}{s}$$

Für den Beispielttest Vitamine und Mineralstoffe betrug der Mittelwert der Rohwerte $\bar{x} = 12$ und die Standardabweichung $s = 3$. Für den Rohwert 15 ergibt sich:

$$z = \frac{15 - 12}{3} = \frac{3}{3} = +1$$

Entsprechend gilt für den Rohwert 9:

$$z = \frac{9 - 12}{3} = \frac{-3}{3} = -1$$

Ein Schüler, der 12 Rohwerte erzielte, erhält den z-Wert:

$$z = \frac{12 - 12}{3} = \frac{0}{3} = 0$$

Vorausgesetzt die Rohwerte sind annähernd normal verteilt, lassen sich so die Rohwerte eines Testes in eine Standardnormalverteilung überführen (vgl. Abbildung 11). In Tabelle 15 sind für alle Rohwerte im Test Vit-

Rohwert	z-Wert	Häufigkeit (f)	$z = \frac{x - \bar{x}}{s}$ $\bar{x} = 12$ $s = 3$
18	2,0	1	
17	1,7	1	
16	1,3	2	
15	1,0	2	
14	0,7	3	
13	0,3	5	
12	0,0	2	
11	-0,3	5	
10	-0,7	3	
9	-1,0	0	
8	-1,3	4	
7	-1,7	1	
6	-2,0	1	
Summe		30	

Tab. 15:

Umsetzung der Rohwerte des Testes Vitamine und Mineralstoffe in Standardwerte z .

amine und Mineralstoffe die entsprechenden z-Werte berechnet. In der Spalte rechts neben den z-Werten ist die Häufigkeit vermerkt, mit der die z-Werte bzw. die zugehörigen Rohwerte vertreten sind. Vergleichen Sie hierzu Tabelle 3, Seite 17.

Ein Schüler, der in diesem Test 8 Rohwerte erzielte, erhält einen z-Wert von $z = -1,3$. Das heißt, daß er das 1,3fache der Standardabweichung unter dem Mittelwert liegt. Sein Prozentrang liegt unter $PR = 16$, da der Prozentrang für $z = -1$ erst 16 ist (vgl. Abb. 11).

Die ungefähre Größe der Standardabweichung einer Rohwertverteilung läßt sich schätzen, sofern die Rohwerte annähernd normal verteilt sind. Für den Test Vitamine und Mineralstoffe wurde eine Standardabweichung von $s = 3,00$ errechnet (S. 38). Hat der errechnete Wert eine Größenordnung, die sinnvoll ist? Im Bereich von $\bar{x}$ bis $\bar{x} + 1 \cdot s$ müssen 34 % aller Werte der Verteilung liegen. Dies sind ca. 1/3 aller Werte. 30 Schüler wurden getestet. 1/3 davon sind 10. Wenn sie in Tabelle 15, vom Mittelwert 12 ausgehend, in der Spalte Häufigkeit 10 Besetzungshäufigkeiten nach oben abzählen, gelangen Sie zum Rohwert 15 oder 14, je nach dem ob die Besetzungshäufigkeit des Mittelwertes mitgezählt wird oder nicht. Die Differenz zwischen dem Rohwert 15 bzw. 14 und dem Mittelwert 12 ist 3 bzw. 2. Die errechnete Standardabweichung wird also $s = 3$ oder $s = 2$ betragen. Die tatsächlich errechnete Standardabweichung von $s = 3,00$ hat eine sinnvolle Größenordnung.

Ähnlich läßt sich auch qualitativ überprüfen, ob eine aktuelle Verteilung normal verteilt ist. In der Standardnormalverteilung oder z-Verteilung liegen 68 %, bzw. ca. 2/3 aller Werte im Bereich von $z = -1$ bis $z = +1$. Links und rechts dieses Bereiches liegen 16 %, bzw. ca. 1/6 der Werte. In Tabelle 15 kann ausgezählt werden, daß zwischen $z = -1$ und $z = +1$, bzw. zwischen den Rohwerten 9 und 15, 20 Rohwerte liegen. (Die Gren-

zen des Intervalls sind hier mitgezählt.) Dies sind $\frac{20}{30} = 0,66$. Das heißt

66 % aller Werte liegen in diesem Intervall. 66 % kommt nahe an den gewünschten Wert von 68 % heran. 4 Werte liegen rechts dieses Bereiches. Dies sind $4/30 = 0,13$, d.h. 13 %. In einer Normalverteilung wären 16 % gefordert. 6 Werte liegen links des Bereiches, d.h. unter $z = -1$. $6/30 = 0,20$. Dies sind 20 %. In einer Normalverteilung wären hier wiederum 16 % der Werte zu erwarten (vgl. Abbildung 10 und 11). – Die im ersten Testdurchlauf für den Test Vitamine und Mineralstoffe erreichten Rohwerte bei 30 Berufsfachschülern können somit angenähert als normal verteilt angesehen werden. Zwischen $\bar{x} \pm 1 \cdot s$ sowie links und rechts dieses Bereiches fallen etwa so viele Werte, wie es von einer normalen Verteilung erwartet wird.

Bedeutung der Standardwerte z

Mit Standardwerten lassen sich die Leistungen zweier Tests vergleichen, sofern die Tests unterschiedliche Standardabweichungen und Variationsweiten von Rohwerten aufweisen. So möge z. B. ein Schüler Wilfried K. in einem Mathematiktest 7,6 Rohwerte erreichen. In diesem Test liegt der Mittelwert bei $\bar{x} = 6,0$ und die Standardabweichung beträgt $s = 1,6$. Der gleiche Schüler erreichte in einem Physiktest 80 Rohwerte. Bei diesem Test lag das Mittel bei $\bar{x} = 50$ und die Standardabweichung betrug $s = 15$. In welchen der beiden Tests schnitt Wilfried K. im Vergleich zu seinen Mitschülern besser ab? Über den Standardwert z läßt sich dies leicht ablesen. Für den Mathematiktest beträgt der z -Wert für Wilfried K.:

$$z_{\text{Mathematik}} = \frac{7,6 - 6,0}{1,6} = \frac{1,6}{1,6} = 1$$

und für den Physiktest:

$$z_{\text{Physik}} = \frac{80 - 50}{15} = \frac{30}{15} = 2$$

Im Physiktest beträgt der z -Wert $z = 2$. Hier liegt Wilfried K. mit seiner Leistung um 2 Standardabweichungen über dem Mittelwert ($PR = 98$), während er mit $z = 1$ im Mathematiktest nur eine Standardabweichung über dem Mittel liegt ($PR = 84$). Im Mathematiktest ist seine Leistung relativ gesehen geringer.

Übung 6:

In einem Test mit $\bar{x} = 62$ und $s = 5$ erreicht ein Schüler a einen Rohwert von 70, Schüler b einen Rohwert von 64. Um wieviel ist die Leistung des Schülers a besser als die des Schülers b, ausgedrückt in z -Werten oder Standardabweichungen?

Die Lösung ist auf Seite 58 angegeben.

Lösung der Übungsaufgabe 6:

$$z_a = \frac{70 - 62}{5} = \frac{8}{5} = 1,6$$

$$z_b = \frac{64 - 62}{5} = \frac{2}{5} = 0,4$$

Schüler a erreicht einen z-Wert von $z = 1,6$ (Prozentrang zwischen PR = 84 und PR = 98), während Schüler b einen z-Wert von $z = 0,4$ erreicht (PR von ca. 65, vgl. Abb. 11). Die z-Wertdifferenz zwischen beiden Schülern beträgt 1,2. Das heißt, daß Schüler a das 1,2fache der Standardabweichung besser ist als Schüler b.

Die Umsetzung der Rohwerte in Standardwerte z verändert nicht die ursprüngliche Form der Rohwertverteilung. Schiefe Verteilungen, seien sie links- oder rechtsschief (vgl. S. 29), bleiben schief; symmetrische Verteilungen bleiben symmetrisch. Man sagt in diesem Zusammenhang auch, die Transformation der Rohwerte in Standardwerte z ist geradlinig oder, mehr fachgemäß ausgedrückt, die Transformation ist linear.

5.3 Transformation der Standardwerte z in Standardwertskalen (IQ, C, T u. Z)

Standardwerte z haben zwei Nachteile. (1) Es treten negative Werte auf. Das Vorzeichen muß stets beachtet werden. (2) Der z-Wert ist eine gebrochene Zahl; um ausreichend präzise zu sein, muß der z-Wert auf zumindest eine Stelle hinter dem Komma angegeben werden.

Um nicht mehr mit negativen und gebrochenen Zahlen zu arbeiten, kann man die Standardwerte z in andere Standardwerte transformieren.

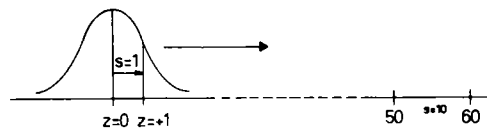


Abb. 12:

Grafische Verdeutlichung der Transformation von Standardwerten z in eine T-Skala: $T = 50 + 10 \cdot z$

Abbildung 12 veranschaulicht solch eine Transformation. Die z-Verteilung wird so nach rechts verschoben, daß ihre senkrechte Achse in $z = 0$ beim Wert 50 einrastet. Die Verteilung ist damit weit aus dem Bereich der negativen Zahlen herausgeschoben. Um Kommastellen bei den neuen Standardwerten zu vermeiden, wird die Standardabweichung der z-Verteilung ($s = 1$) mit dem Faktor 10 multipliziert. $z = 1$ (eine Standardabweichung rechts vom Mittelwert $z = 0$) fällt dann auf den Wert 60. Bei

dieser Transformation gehen die z-Werte in sog. T-Werte über. Die T-Skala ist nach Terman, einem Testtheoretiker, benannt. Formelmäßig lautet die T-Skala:

$$T = 50 + 10 \cdot z$$

Die Grundformel für die Transformation der Standardwerte z in andere Standardwerte lautet:

Definition 16:

Transformation der Standardwerte z

Standardwert = $M + S \cdot z$

mit M = gewünschter Mittelwert,

S = gewünschte Standardabweichung,

$$z = \frac{\text{Rohwert minus Mittelwert der Rohwerte}}{\text{Standardabweichung der Rohwerte}}$$

Bekannt ist die IQ-Skala, die einen Mittelwert von $M = 100$ und eine Standardabweichung von $S = 15$ hat:

$$IQ = 100 + 15 \cdot z$$

Verwendet wird auch die Centilskala, Mittelwert $M = 5$, Standardabweichung $S = 2$:

$$C = 5 + 2 \cdot z$$

Zu nennen wäre weiterhin die („groß“)Z-Skala mit $M = 100$ und $S = 10$:

$$Z = 100 + 10 \cdot z$$

Letztere Skala wird widersinnig auch allein mit Standardwertskala (SW) bezeichnet, obwohl auch die bisher vorgestellten z-, IQ-, C- und T-Werte bereits Standardwertskalen sind. Ebenso werden die („groß“) Z-Werte oft allein als Standardwerte bezeichnet. Hier im folgenden aber soll nur von Z-Werten gesprochen werden.

IQ- und C-Skalen sind in der Regel den Intelligenztests vorbehalten. Bei standardisierten Schulleistungstests oder vom Lehrer selbst erstellten informellen Leistungstests werden die Rohwerte in der Regel in der T-Skala ausgedrückt. So sind in Tabelle 16 die z-Werte für den Test Vitamine und Mineralstoffe (Tabelle 15) in T-Werte umgeformt worden.

Ein Schüler, der einen Rohwert von $RW = 8$ im Test Vitamine und Mineralstoffe erreichte, erhält einen T-Wert von 37. Seine Leistung liegt mehr als eine Standardabweichung (13 T-Werte) unter dem Mittelwert von $T = 50$. (Die Standardabweichung in der T-Skala beträgt 10 T-Werte.)

Bei der Transformation der Rohwerte in eine Standardwertskala T sind zwei Arten von Standardabweichungen immer auseinander zu halten. Zur

Rohwert	Häufigkeit	z-Wert	T-Wert
18	1	2,0	70
17	1	1,7	67
16	2	1,3	63 = 50 + 10 · 1,3
15	2	1,0	60
14	3	0,7	57
13	5	0,3	53
12	2	0,0	50
11	5	-0,3	47
10	3	-0,7	43
9	0	-1,0	40
8	4	-1,3	37
7	1	-1,7	33
6	1	-2,0	30

Tab. 16:

Umsetzung der Rohwerte des Testes Vitamine und Mineralstoffe in T-Werte.

Bestimmung des z-Wertes wird die Standardabweichung der Rohwertverteilung, „klein“ s , verwendet. Bei der Transformation der z-Werte wird wieder eine Standardabweichung benutzt. Hier handelt es sich aber um die gewünschte Standardabweichung „groß“ S der Skala (vgl. Def. 16!!). „Groß“ S ist ein Faktor, mit der die Standardabweichung „klein“ s gestreckt wird!

Die Transformation der Standardwerte z in andere Standardwerte, wie z. B. in T-Werte, ist linear. Schiefen der Rohwertverteilung gehen in gleiche Schiefen bei der z -Verteilung und in gleiche Schiefen bei der T-Verteilung über.

Neben den T-, IQ-, Z- und C-Skalen lassen sich beliebig viele andere konstruieren. Man kann z. B. die Rohwerte eines informellen Leistungstests, den man selber erstellt hat, in eine selbstentwickelte Skala umsetzen, z. B. in eine Standardwertskala $25 + 5 \cdot z$, oder $500 + 100 \cdot z$. Es empfiehlt sich jedoch, bei einem Leistungstest, da es bereits so viele Skalen gibt, konventionsgemäß die T-Skala zu verwenden.

5.4 Exkurs: Der ältere IQ-Begriff

Der Intelligenz-Quotient (IQ) ist definiert als Quotient aus Intelligenzalter und Lebensalter, multipliziert mit 100, um gebrochene Zahlen zu vermeiden.

$$IQ = \frac{\text{Intelligenzalter}}{\text{Lebensalter}} \cdot 100$$

Das Intelligenzalter eines Probanden in einem Test entspricht dem Jahrgang, dessen Durchschnittsrohwert seinem Rohwert gleicht. Erreicht ein Proband A – 9,0 Jahre alt – in einem Intelligenztest 120 Rohwerte und entsprechend 120 Rohwerte dem Wert, den alle 10,5-jährigen der Eich-

stichprobe im Schnitt in diesem Test erreichten, so hat Proband A ein Intelligenzalter von 10,5 Jahren. Sein IQ beträgt demnach:

$$IQ = \frac{10,5}{9,0} \cdot 100 = 117.$$

Diese Berechnung des Intelligenzquotienten hat zwei Nachteile:

1) Innerhalb der verschiedenen Altersklassen kann die Standardabweichung der Rohwerte unterschiedlich groß sein. Da die Standardabweichung bei der Transformation in den Intelligenzquotienten nicht mit eingeht, werden diese Unterschiede nicht berücksichtigt.

2) Der Quotient wird widersinnig, wenn damit – ohne weitere Kunstgriffe – die Intelligenzleistung Erwachsener ausgedrückt werden soll. Hat ein 40jähriger Mann das Intelligenzalter eines 20jährigen, so beträgt sein IQ:

$$IQ = \frac{20}{40} \cdot 100 = 50!$$

Aufgrund dieser Nachteile wurde der hier bereits vorgestellte Abweichungs-IQ ($IQ = 100 + 15 \cdot z$) eingeführt, der eigentlich kein Quotient mehr ist. Der ältere IQ-Begriff dürfte heute wohl weniger verwendet werden.

5.5 Umrechnung von Standardwertskalen

Wichtig ist es, daß man in verschiedenen Skalen „denken“ kann. Ein T-Wert von 60 entspricht einem IQ-Wert von 115. Beide liegen eine Standardabweichung über dem Mittelwert. Gleichfalls eine Standardabweichung über dem Mittelwert liegen $C = 7$ und $Z = 110$. Ein T-Wert von 40 liegt eine Standardabweichung unter dem Mittelwert. Dies entspricht einem IQ-Wert von 85, einem Centilwert von 3 und einem Z-Wert von 90.

Welchem Centilwert C entspricht der IQ-Wert 117? Über die Skalen-Gleichungen läßt sich ausrechnen:

(Es ist $IQ = 100 + 15 \cdot z$ u. $C = 5 + 2 \cdot z$)

$$117 = 100 + 15 \cdot z, \text{ daraus folgt für } z:$$

$$117 - 100 = 15 \cdot z$$

$$\frac{117 - 100}{15} = z$$

$$z = 1,13$$

$$z = 1,13 \text{ einsetzen in } C = 5 + 2 \cdot z, \text{ daraus folgt für } C:$$

$$C = 5 + 2 \cdot 1,13$$

$$C = 7,26$$

Dem IQ-Wert von 117 entspricht ein Centilwert C von 7,26.

Sollen mehrfach IQ-Werte in Centilwerte umgerechnet werden, empfiehlt sich eine „programmierte Berechnung“ (Tabelle 16a). Einfache technisch-wissenschaftliche Taschenrechner sind bereits programmierbar. Sie gestatten es, über eine Lern-Taste Formeln bzw. Rechengänge für automatische Wiederholungsberechnungen aufzunehmen.

Die Programmierung ist denkbar einfach. Wie aus Tabelle 16a zu ersehen ist, lassen sich die Programmschritte zur Umrechnung eines IQ-Wertes in einen Centilwert in gleicher Form eintasten, wie oben die Umrechnung schrittweise „zu Fuß“ vorgenommen wurde. Nach der Rückstellung für die 1. Berechnung am Ende des Programms wird ein beliebiger IQ-Wert eingetastet (hier 117). Durch Drücken der Programmstart-Taste

Tasten	Bemerkungen/Anzeige
2nd CA	Gesamtlöschung
2nd Fix 2	Festkomma auf 2 Dezimalstellen
2nd Lrn	Eingabe des Lern-Modus
-	
100	
=	
÷	
15	Eintasten der Programm- schritte
=	
x	
2	
+	
5	
=	
2nd R/S	Programmstopp/Anzeige des Resultats
2nd Rst	Programmrückstellung
2nd Lrn	Abstellen des Lern-Modus
2nd Rst	Rückstellung für die 1. Berechnung
117	Eingabe eines beliebigen IQ - Wertes
2nd R/S	Programmstart
	7,26 (Ergebnis: Centilwert entsprechend zum IQ von 117)

Tab. 16a:

Programmierte Umrechnungen von IQ-Werten in Centilwerte anhand eines elektronischen Taschenrechners (Texas Instruments TI – 51 – III).

Tastenerklärung:

2nd	(second) Zugang zur Tastaturebene 2
CA	Clear all
Fix	Festkommavorgabe
Lrn	Learn-Modus
R/S	Run/Stop
Rst	Reset (Rückstellen)
$\div$	"geteilt durch"
$\times$	"mal"

erhalten Sie sofort den entsprechenden Centilwert von 7,26. Es läßt sich nun jeder andere IQ-Wert eintasten, für den man nach Tasten des Programmstartes sofort den entsprechenden Centilwert auf zwei Dezimalstellen genau erhält. – In ähnlicher Form sind Umrechnungen von Centilwerten in IQ-Werte oder von T-Werten in Z-Werte sowie auch andere Umrechnungen programmierbar.

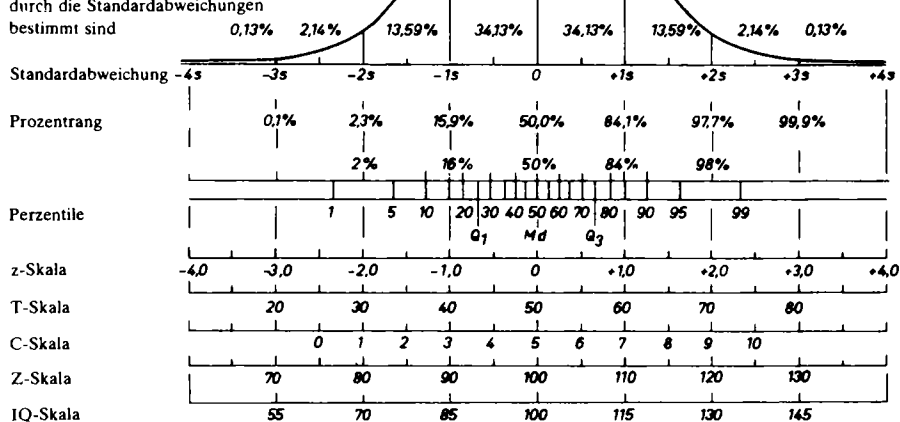
Übung 7:

Berechnen Sie bitte schrittweise nach S. 61

- a) zum T-Wert 43 den entsprechenden IQ-Wert und
 - b) zum Centilwert 4,56 den entsprechenden Z-Wert.
- Die Lösungen sind auf S.65 angegeben.

In Tabelle 17 sind die verschiedenen Standardwertskalen zusammengestellt. Mit angegeben sind die Prozentränge oder Perzentile, die mit den einzelnen Standardwerten korrespondieren. Oberhalb der Skalen ist die Standardnormalverteilung ($N(0,1)$) abgebildet. Die Auftretenswahrscheinlichkeit von Werten in bestimmten Intervallen einer Standardskala läßt sich so ablesen. Ein T-Wert im Intervall von 30 bis 40 hat eine Auftretenswahrscheinlichkeit von 0,1359. Das heißt, daß in einer normalen Verteilung 13,59% aller Werte oder Fälle in das Intervall von $T = 30$ bis $T = 40$ fallen. Aus Tabelle 17 läßt sich weiter ablesen, daß ein IQ-Wert von 85 einem Z-Wert von 90, einem C-Wert von 3, einem z-Wert von $-1,0$ und einem Prozentrang von 16 entspricht. Zwischenwerte lassen sich aus Tabelle 17 nur ungefähr ablesen. Hier sei auf die Ausrechnung verwiesen, wie sie oben durchgeführt worden ist. Eine Hilfe bietet auch Tabelle 18.

Prozent der Fälle in den
Intervallabschnitten, die
durch die Standardabweichungen
bestimmt sind



Tab. 17:
Standardnormalverteilung, Prozentränge und Standardwerte (Grafik: K. Ingen-
kamp (Hrsg.), Tests in der Schulpraxis, Weinheim 1971, S. 153).

Lösung zur Übungsaufgabe 7:

$$a) T = 50 + 10 z, 43 = 50 + 10 z \Rightarrow z = -0,7$$

$$IQ = 100 + 15 z = 100 + 15 (-0,7) = \underline{89,5}.$$

$$b) C = 5 + 2 z, 4,56 = 5 + 2 z \Rightarrow z = -0,22$$

$$Z = 100 + 10 z = 100 + 10 (-0,22) = \underline{97,8}.$$

Tabelle 18 gibt ganzzahlige Prozentränge und ihre Äquivalente in den verschiedenen Standardwerten wieder. Ein T-Wert von 66 entspricht einem IQ von 124, einem C-Wert von 8, einem Z-Wert von 116 und einem Prozentrang von 95. Der Centilwert ist hier ganzzahlig angegeben.

Prozent-rang	T-Wert	IQ-Wert	C-Wert	Z-Wert	Prozent-rang	T-Wert	IQ-Wert	C-Wert	Z-Wert
100	80	145	11	130	50	50	100	5	100
100	79	144	11	129	46	49	99	5	99
100	78	142	11	128	42	48	97	5	98
100	77	141	10	127	38	47	96	4	97
100	76	139	10	126	34	46	94	4	96
99	75	138	10	124	31	45	93	4	95
99	74	136	10	124	27	44	91	4	94
99	73	135	10	123	24	43	90	4	93
99	72	133	9	122	21	42	88	3	92
98	71	132	9	121	18	41	87	3	91
98	70	130	9	120	16	40	85	3	90
97	69	129	9	119	14	39	84	3	89
96	68	127	9	118	12	38	82	3	88
96	67	126	8	117	10	37	81	2	87
95	66	124	8	116	8	36	79	2	86
93	65	123	8	115	7	35	78	2	85
92	64	121	8	114	5	34	76	2	84
90	63	120	8	113	4	33	75	2	83
88	62	118	7	112	4	32	73	1	83
86	61	117	7	111	3	31	72	1	82
84	60	115	7	110	2	30	70	1	80
82	59	114	7	109	2	29	69	1	79
79	58	112	7	108	1	28	67	1	78
76	57	111	6	107	1	27	66	0	77
73	56	109	6	106	1	26	64	0	76
69	55	108	6	105	1	25	63	0	75
66	54	106	6	104	0	24	61	0	74
62	53	105	6	103	0	23	60	0	73
58	52	103	5	102	0	22	58	-1	72
54	51	102	5	101	0	21	57	-1	71
50	50	100	5	100	0	20	55	-1	70

Tab. 18:

Prozentrang und Standardwerte T, IQ, C u. Z.

5.6 Exkurs: Nicht lineare Transformationen

In Tabelle 18, S. 65, ist neben den T-, IQ-, C- und Z-Werten der Prozentrang mit aufgeführt. Angenommen, vier Personen wurden einem Intelligenztest unterzogen. Person A erreichte einen IQ von 130, Person B er-

zielte einen IQ von 115. Die Differenz beträgt 15 IQ-Einheiten. Die entsprechende Prozentrangdifferenz beträgt $98 - 84 = 14$ (s. Tab. 18). Person C erreichte einen IQ von 100, während Person D bei einem IQ von 85 liegt. Die Leistungsdifferenz beträgt 15 IQ-Einheiten, wie zwischen Person A und B. Die entsprechende Prozentrangdifferenz zwischen Person C und D ist jedoch größer als die Differenz zwischen Person A und B. Sie beträgt $50 - 16 = 34$ Prozentrangplätze (s. Tab. 18). Bei gleichbleibender IQ-Differenz können sich unterschiedliche Prozentrangdifferenzen ergeben. Aus Tabelle 17 können Sie ersehen, daß die Prozentrangskala in der Mitte gestaucht und an den Enden gestreckt ist. Die Prozentrangskala ist ungeradlinig („nicht linear“). Man nennt die Prozentrangskala auch eine Flächen-transformierende Skala. Sie gibt die Flächenanteile der Normalverteilung wieder (s. Tab. 17). IQ-, T-, C- und Z-Skalen sind im Gegensatz dazu linear transformierende Skalen.

Mit der Prozentrangskala lassen sich anormal verteilte Rohwerte in normal verteilte Werte transformieren. Anormale Verteilungen lassen sich normalisieren, indem von diesen Verteilungen zunächst die Prozentränge berechnet werden, also eine Flächentransformation vorgenommen wird, und dann (nach Tabelle 18) diese Prozentränge in Standardwert-Äquivalente umgewandelt werden.

Abb. 13 gibt einen solchen Normalisierungsvorgang grafisch wieder. Von der anormalen Verteilung der Rohwerte in einem Test mit maximal 30 erreichbaren Rohwerten werden die entsprechenden Prozentränge be-

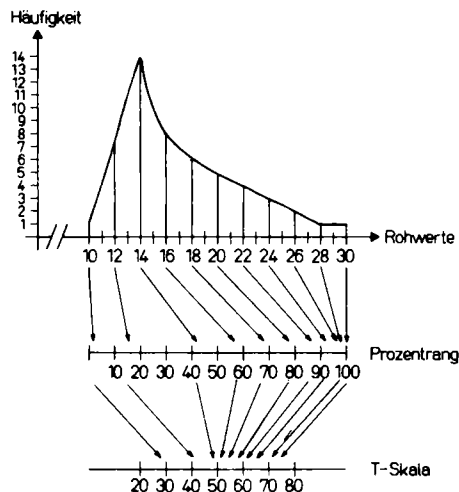


Abb. 13:
Transformation einer Rohwertverteilung in eine Standardnormalverteilung über Prozentränge (nichtlineare Transformation: Flächentransformation).

rechnet (s. Def. 14, S. 45). Von den Prozenträngen ausgehend lassen sich dann Standardwert-Äquivalente angeben (Tab. 18). In Abb. 13 ist die T-Skala verwandt worden.

Bei der Transformation der Rohwerte eines Testes bestehen zwei Möglichkeiten:

- Die Rohwerte werden direkt – wie in vorherigen Kapiteln besprochen – in Standardwerte z , T , C oder Z umgewandelt. Diese Transformation ist linear. Asymmetrien der Rohwertverteilung bleiben in den Standardwertverteilungen erhalten.
- Die Rohwerte werden über den Umweg über Prozentränge in Standardwerte z , T , C , Z umgewandelt. Diese Transformation ist nicht linear. Asymmetrien der Rohwertverteilung werden geglättet, so daß eine Standardwertverteilung entsteht, die normal verteilt ist.

Bei der Entwicklung informeller Leistungstests dürfte es sich empfehlen, die Rohwerte nicht über die Prozentränge zu normalisieren. Solange die Rohwerte linear über z in Standardwerte T , Z oder C transformiert werden, bleibt die mögliche Unzulänglichkeit des entwickelten Testes („krumme Verteilung“) erhalten. Die Notwendigkeit für eine Testrevision kann durch eine Normalisierung der Standardwertverteilung verwischt werden. Nach der Glättung einer „krummen“ Rohwertverteilung in eine Normalverteilung entsteht der Eindruck, man habe es nun mit einem gut differenzierenden Meßverfahren zu tun⁸.

5.7 Stanine, Stasix und Sten

Stanine

Neben den bisher besprochenen Standardwertskalen ist noch die Stanine-Skala zu nennen (*Standard scores with nine categories*.) Bei der Stanine-Skala wird eine Normalverteilung, ausgehend vom Mittelwert, nach links und rechts längs der Horizontalen in neun gleich große Abschnitte aufgeteilt (Abbildung 14).

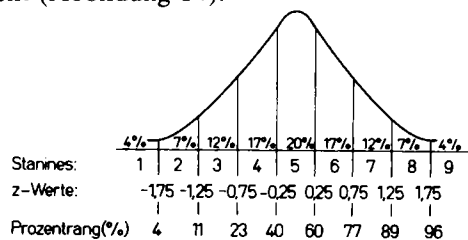


Abb. 14:

Beziehung zwischen der Stanine-Skala, der z -Skala und der Prozentrangskala in der Standardnormalverteilung.

Der mittlere Stanine-Abschnitt 5 reicht von $z = -0,25$ bis $z = +0,25$; er umfaßt also eine halbe Standardabweichung der Standardnormalverteilung. Von diesem mittleren Abschnitt ausgehend ist nun die Standardnormalverteilung nach links und rechts in gleich große Abschnitte aufgeteilt. In Tabelle 19 sind aus Abbildung 14 die Stanine Werte, ihre Intervallgrenzen z und die Häufigkeit des Vorkommens dieser Intervalle herausgeschrieben. Zusätzlich sind die Prozentränge angegeben; es wurde auch versucht, den numerischen Wert eines Stanines verbal zu beschreiben⁹.

Für die Überführung der Rohwerte eines Testes in Stanines bieten sich zwei Möglichkeiten an:

- a) von den Rohwerten über die z -Werte zu den Stanines oder
- b) von den Rohwerten über Prozentränge zu den Stanines.

Zu a):

Sofern über die (nicht über Prozentränge ermittelten) z -Werte die Stanines für eine Rohwertverteilung ausgerechnet werden, wird eine u.U. schiefe oder unregelmäßige Rohwertverteilung dadurch nicht normalisiert. Tabelle 20 zeigt das Vorgehen für die Rohwerte des Beispieltests Vitamine und Mineralstoffe. In Tabelle 15, S. 55, sind bereits die z -Werte für die Rohwerte dieses Testes berechnet worden. Gemäß den in Tabelle 19 oder in Abbildung 14 angegebenen z -Wert-Intervallen sind zu diesen z -Werten in Tabelle 20 die entsprechenden Stanines angegeben.

Stanine	Verbale Beschreibung	Häufigkeit (%)	Prozentrang(PR) (%)	Intervall in z -Werten
9	Außerst hoch	4	100	1,75 bis $+\infty$
8	Sehr hoch	7	96	1,25 bis 1,75
7	Hoch	12	89	0,75 bis 1,25
6	Oberer Durchschnitt	17	77	0,25 bis 0,75
5	Durchschnitt	20	60	-0,25 bis 0,25
4	Unterer Durchschnitt	17	40	-0,75 bis -0,25
3	Niedrig	12	23	-1,25 bis -0,75
2	Sehr niedrig	7	11	-1,75 bis -1,25
1	Außerst niedrig	4	4	$-\infty$ bis -1,75

Tab. 19:

Verbale Beschreibung, Häufigkeit, Prozentränge und Intervalle von Stanines.

Zu b):

Sofern die Rohwerte normalisiert werden sollen, wird bei der Berechnung der Stanines von dem Prozentrang der Rohwerte ausgegangen. Tabelle 21 zeigt das Vorgehen. In Tabelle 13, S. 46, sind die Prozentränge für die Rohwerte des Testes Vitamine und Mineralstoffe berechnet worden. Gemäß

Rohwert	z-Wert	Stanine
18	2,0	9
17	1,7	8
16	1,3	8
15	1,0	7
14	0,7	6
13	0,3	6
12	0,0	5
11	-0,3	4
10	-0,7	3
9	-1,0	3
8	-1,3	2
7	-1,7	2
6	-2,0	1

Tab. 20:

Überführung der Rohwerte des Testes Vitamine und Mineralstoffe ($\bar{x} = 12, s = 3$) in lineare Stanines; nach Abbildung 14.

den in Tabelle 19 oder in Abbildung 14 angegebenen Prozenträngen lassen sich nun zu den Prozenträngen der Rohwerte in Tabelle 21 die Stanines angeben. Der Rohwert 10 hat z. B. den Prozentrang 30. Dieser Prozentrang liegt zwischen den Prozenträngen 23 und 40 (vgl. Abbildung 14 oder Tabelle 19). Dies sind die Prozentrangsgrenzen von Stanine 4. Dem Rohwert 10 wird der Stanine-Wert 4 zugeordnet.

Rohwert (RW)	Häufigkeit (f)	cum _f	Prozentrang (PR) (%)	Stanine
18	1	30	100	9
17	1	29	96,66	9
16	2	28	93,33	8
15	2	26	86,66	7
14	3	24	80	7
13	5	21	70	6
12	2	16	53,33	5
11	5	14	46,66	5
10	3	9	30	4
9	0	6	20	3
8	4	6	20	3
7	1	2	6,66	2
6	1	1	3,33	1

N = 30

Tab. 21:

Überführung der Rohwerte des Testes Vitamine und Mineralstoffe in normalisierte Stanines.

Ein Vergleich der Tabellen 20 und 21 zeigt, daß sich die Stanine-Zuordnungen zu den Rohwerten in den beiden Tabellen kaum unterscheiden. Bereits früher war für den Test Vitamine und Mineralstoffe festgestellt worden, daß die Rohwerte sich in etwa normal verteilen (S. 56). Demzufolge kann auch kein großer Unterschied mehr zwischen linear und normalisiert ermittelten Stanines bestehen.

Zu dem Problem der Normalisierung einer Rohwertverteilung wurde bereits oben Stellung genommen (S. 67): Bei der Konstruktion informeller Leistungstests sollten die Rohwerte immer linear transformiert werden. Das bedeutet für eine Transformation von Rohwerten in Stanines, daß nach Verfahren a (Tabelle 20) vorzugehen ist, indem die Rohwerte über die (nicht normalisierten) z-Werte in Stanines überführt werden.

Stanines lassen nur ganzzahlige, ungebrochene Werte zu. Personen, die unterschiedliche Rohwerte in einem Test erreicht haben, können dem gleichen Stanine-Wert zugeordnet werden. Erzielten z. B. zwei Schüler in dem Test Vitamine und Mineralstoffe die Rohwerte 13 und 14, so erhalten beide den Stanine-Wert 6 (Tabelle 20). Das heißt, durch das Stanine-Transformationsverfahren geht Information verloren.

Die Skala kann die Rohwerte nur in grobe Klassen einteilen. Sobald in einem Test mehr als 9 Rohwerte erreicht werden, und in den meisten Tests ist dies ja der Fall, wird sich ein Informationsverlust bezüglich der Differenzierungsleistung des Testes einstellen. Je mehr Rohwerte in einem Test vergeben werden können, desto mehr Rohwerte fallen dann auch unter einen Stanine-Wert. Dies ist der Preis, der gezahlt werden muß, wenn man mit einer handlichen Skala arbeiten will, die ganzzahlig ist und nur 9 Einheiten umfaßt.

Stasix

So wie sich eine Standardskala mit 9 Einheiten konstruieren läßt, kann auch eine mit z. B. 6 Einheiten gebildet werden: Stasix (*Standard scores with six categories*).

Die Stasix Skala sieht sechs Abschnitte bei einem mittleren 50%-Bereich links und rechts des Mittelwertes vor¹⁰. Ausgehend vom Mittelwert $z = 0$ der Standardnormalverteilung wird ein Abschnitt so nach rechts abgetragen, daß ihm 25 % der Normalverteilung entsprechen ($z = 0$ bis $z = 0,67$). Um den gleichen Abstand erweitert, ergibt sich $z = 0,67 + 0,67 =$

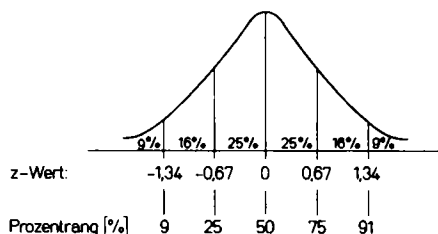


Abb. 15:

Beziehung zwischen der Stasix-Skala, der z-Skala und den Prozenträngen.

1,3 (25 % + 16 % = 41 %). Entsprechend wird von $z = 0$ nach links hin die Standardnormalverteilung aufgeteilt. Es ergeben sich sechs Abschnitte, einschließlich der offenen Enden (Abbildung 15).

Sten

Verwendung findet auch eine Standardskala mit zehn Einteilungen: Sten (Standard scores with *ten* categories). Ausgehend vom Mittelwert $z = 0$ wird die Standardnormalverteilung in halben Standardwerteinheiten nach links und rechts aufgeteilt (Tab. 21a). Auf einen Sten-Wert von 7 entfallen in einer Normalverteilung 15 % aller Werte. Der Sten-Wert 7 liegt zwischen $z = 0,5$ und $z = 1$ bzw. zwischen den Prozentranggrenzen 69 und 84.

Prozent der Fälle in den Abschnitten	2%	5%	9%	15%	19%	19%	15%	9%	5%	2%
z - Wert	-2	-1,5	-1,0	-0,5	0	0,5	1	1,5	2	
Stens	1	2	3	4	5	6	7	8	9	10
Prozentrang (%)	2	7	16	31	50	69	84	93	98	

Tab. 21a:
Beziehung zwischen der z-Skala, Sten-Skala und den Prozenträngen.

Neben der Stanine der Stasix und der Sten Einteilung ließe sich nun auch jede andere entwerfen: z.B. eine „Stafour“ oder eine „Staseven“. Bei geradzahligen Einteilungen ist vom Mittelwert $z = 0$ nach links und rechts auszugehen, wie bei der Stasix oder Sten Einteilung (Abb. 15, Tab. 21a). Bei ungeradzahligen Einteilungen ist ein mittlerer Ausgangs-Abschnitt um den Mittelwert $z = 0$ anzulegen, wie bei der Stanine-Skala (Abb. 14). Es empfiehlt sich jedoch aufgrund der Vielzahl der bereits bestehenden Skalen, nicht noch neue Skalen hinzuzufügen.

Übung 8:

Der Schüler Michael K. erreicht in einem Schulleistungstest ($\bar{x} = 18$, $s = 3,5$) einen Rohwert von 20. Welchen linear transformierten Stanine-Wert erhält Michael K (s. Abb. 14)? Die Lösung ist auf der folgenden Seite angegeben.

Lösung der Übungsaufgabe 8:

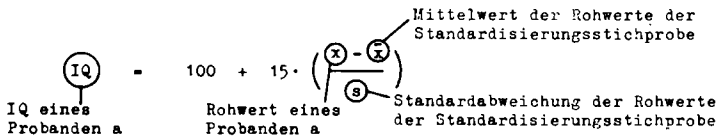
$$z = \frac{x - \bar{x}}{s} = \frac{20 - 18}{3,5} = 0,57$$

$$z = 0,57 \text{ \# Stanine 6 (Abb. 14)}$$

6. Normen

Standardisierte, auf dem Markt käufliche Tests, zeichnen sich u. a. dadurch aus, daß die Leistung eines Probanden mit der Leistung einer repräsentativen Gruppe von Probanden verglichen werden kann. Diese repräsentative Gruppe bezeichnet man als die Standardisierungsstichprobe oder als die Eichstichprobe. Die Standardisierungsstichprobe steht für die Population, auf die der Test abstellt. Bei Intelligenztests umfassen Standardisierungsstichproben ein bis mehrere Tausend Personen. Von den Rohwerten, die diese Personen in dem Test erreichen, werden der Mittelwert $\bar{x}$ und die Standardabweichung s berechnet. Diese zwei Werte reichen aus, um die Leistung eines Probanden mit der Leistung der repräsentativen Gruppe von Probanden zu vergleichen.

Es gibt verschiedene Skalen, in welcher der Mittelwert $\bar{x}$ und die Standardabweichung s der Standardisierungsstichprobe eingehen kann, z. B. in die IQ-, C- oder Z-Skala. Für den IQ gilt z. B.:

$$\text{IQ eines Probanden } a = 100 + 15 \cdot \frac{(\text{Rohwert eines Probanden } a - \text{Mittelwert der Rohwerte der Standardisierungsstichprobe})}{\text{Standardabweichung der Rohwerte der Standardisierungsstichprobe}}$$


An einem Intelligenztest soll das bisher Gesagte verdeutlicht werden.

6.1 Tabellarische Normen

Der Grundintelligenztest CFT 3 nach *R. Cattell*¹¹ nimmt für sich in Anspruch, die Intelligenz relativ kulturfrei zu messen (CFT-Culture Fair Test). Mit diesem Test soll eine grundlegende geistige Fähigkeit gemessen werden, die frei von sprach- und vorbildungsgebundenen Einflüssen ist. Der Test besteht aus zwei Teilen. In jedem Teil sind maximal 50 Rohwerte zu erreichen, d. h. zusammen maximal 100 Rohwerte. Der Grundintelligenztest wurde u. a. auch an Berufsschülern standardisiert, bzw. „geeicht“. Tabelle 22 gibt als Auszug aus der Handanweisung dieses Tests die Schulstandardwerte für Berufsschüler (14–18 Jahre) wieder.

Wie Sie den „Anmerkungen zur Standardisierung“ am Fuß der Tabelle 22 entnehmen können, bildeten die Standardisierungsstichprobe $N = 905$ nordbayrische gewerbliche und kaufmännische Berufsschüler. Der arithmetische Mittelwert der Rohwerte, die diese Schüler insgesamt erreichten, beträgt $\bar{x} = 39,4$. Die Standardabweichung liegt bei $s = 9,6$. Aus

Rohpunkte			Standardwerte	
1. Teil	2. Teil	Ges. 1. u. 2. T.	IQ	T
—	3	—	55	20
3	4	13	57	22
—	5	—	60	23
—	6	—	61	24
—	—	14	62	25
4	7	—	63	25
—	—	15	64	26
—	8	—	66	27
5	—	16	67	28
—	9	—	68	28
6	—	17	69	29
—	10	18	70	30
7	—	—	71	30
—	—	19	72	31
—	11	20	73	32
8	—	—	74	32
9	—	21	75	33
—	12	22	76	34
—	—	23	77	34
—	13	24	78	35
10	—	25	79	36
11	14	26	81	37
—	—	27	82	38
12	15	—	83	38
—	—	28	84	39
—	16	29	85	40
—	—	30	86	41
13	17	—	87	41
—	—	31	88	42
—	—	32	89	42
14	18	33	90	43
—	—	34	91	44
15	19	—	92	44
—	—	35	93	45
16	20	36	94	46
—	—	37	96	47
17	21	38	97	48
—	22	39	99	49
—	—	40	100	50

Rohpunkte			Standardwerte	
1. Teil	2. Teil	Ges. 1. u. 2. T.	IQ	T
—	—	40	100	50
18	—	—	101	50
—	23	41	102	51
—	—	42	103	52
19	—	—	104	52
—	24	43	105	53
—	—	44	106	54
—	25	—	107	54
20	—	45	108	55
—	—	46	110	56
21	26	47	111	57
22	—	48	113	58
—	27	—	114	59
—	—	49	115	60
23	—	50	117	61
—	28	—	118	62
24	—	51	120	63
—	29	—	121	64
—	—	52	122	65
25	—	53	124	66
—	30	54	126	67
26	—	55	127	68
—	—	56	129	69
—	31	—	130	70
27	—	—	131	70
—	—	57	132	71
—	32	—	134	72
28	—	58	135	73
29	—	—	137	74
30	33	59	139	75
—	34	60	141	76
31	—	61	142	77
—	35	—	143	78
—	—	62	144	79
—	36	63	145	80

Anmerkung zur Standardisierung:

Die Transformation von Rohwerten in Standardwerte (IQ- und T-Werte) erfolgte für eine repräsentative Stichprobe aus nordbayerischen gewerblichen und kaufmännischen Berufsschulen über alle Schuljahrgänge. Alle Schüler hatten ausschließlich **Volksschulvorbildung** (N = 905).

	1. Teil	2. Teil	Ges. (1. + 2. Teil)
Arithmet. Mittel des Rohwerts	17,4	21,8	39,4
Standardabweichung (RW)	5,0	5,4	9,6
Stichprobenfehler des Mittelwerts			
3 Sigma _M (in Rohwerten)	± 0,5	± 0,5	± 0,9

Tab. 22:

Normtabelle des Grundintelligenztestes CFT 3 für Berufsschüler (R. H. Weiß, Grundintelligenztest Skala 3-CFT 3, Handanweisung für die Durchführung, Auswertung und Interpretation, Braunschweig 1971, S. 46).

diesen zwei Werten läßt sich nun jeder beliebige IQ- oder T-Wert berechnen. Für einen Rohwert von insgesamt 48 ergibt sich ein IQ-Wert von

$$\begin{aligned} IQ &= 100 + 15 \cdot \left(\frac{48 - 39,4}{9,6} \right) \\ &= 100 + 15 \cdot 0,895 \\ &= 100 + 13,4 \\ &= \underline{113,4} \end{aligned}$$

In der rechten Zahlenspalte in Tabelle 19 steht in der Rubrik Rohpunkte: Ges. 1 u. 2. T. neben der Zahl 48 der IQ-Wert 113.

Hat jemand einen Rohwert von 48 in diesem Test erreicht, so hat er im Vergleich zu Berufsschülern einen IQ von 113. Der IQ einer getesteten Person wird immer an der Norm einer vergleichbaren Gruppe von Personen ausgedrückt. Man nennt eine Zahlentabelle wie in Tabelle 22 demzufolge auch „Normentabelle“ oder wie beim Grundintelligenztest CFT 3 „Normsignatur“. – Neben der Normentabelle für Berufsschüler enthält der Grundintelligenztest noch eine Gesamtnormentabelle für 16- bis 19jährige (N = 608), eine für Schüler mit mittlerem Schulabschluß (N = 187), eine für Unterprimaner (N = 1012) und Vergleichswerte (Mittelwert und Standardabweichung) für erfolgreich Studierende aus 10 Hochschulfachrichtungen.

Zusammenfassend gilt:

Normen bilden Vergleichsmaßstäbe, an denen die Leistungen einzelner Personen gemessen werden können. Der Vergleichsmaßstab basiert auf der Leistung einer repräsentativen Gruppe von Personen (Bezugsgruppe). Diese u. U. kompliziert ausgesuchten, aber relevanten Personen bilden die Standardisierungs- oder Eichstichprobe. Es ist dabei nicht die Größe der Standardisierungsstichprobe sondern ihre Repräsentativität wichtig.

Bevor Sie zu Übung 9 übergehen sei folgendes deutlich herausgestellt. In Tabelle 22 wird mit zwei Standardabweichungen gearbeitet. Einmal wird die Standardabweichung der Rohwerte ermittelt. Dies ist im Gegensatz zu Definition 16 (S. 59) das „klein“ s, das in die Bestimmung des z-Wertes eingeht. In die Standardwerte – in Tabelle 22 sind dies die IQ- und T-Werte – geht auch eine Standardabweichung ein. Diese Standardabweichung wurde in Definition 16 mit „groß“ S bezeichnet. S ist die für die Skala erwünschte, fest definierte Standardabweichung, mit der die z-Werte multipliziert werden. Z.B. ist

$$IQ = 100 + \underset{\text{"groß" S}}{15} \cdot z = 100 + 15 \cdot \left(\frac{x - \bar{x}}{\underset{\text{"klein" s}}{9}} \right)$$

Übung 9:

Prüfen Sie bitte die Angaben nach Tabelle 22 rechnerisch nach, daß ein Proband mit einem Rohwert von insgesamt 37 einen T-Wert von 47 erhält. – Die Lösung ist auf der nächsten Seite angegeben.

Das Prüfsystem für Schul- und Bildungsberatung (P-S-B) von W. Horn ist ein Intelligenzmeßverfahren, das die Struktur der Intelligenz eines Probanden zu erfassen verspricht¹². In zehn Untertests wird versucht, die Intelligenzleistung unterschiedlicher Intelligenzfaktoren abzuprüfen. Intelligenzfaktoren sind z.B. die „Denkfähigkeit“, die „Technische Begabung“, die „Wortflüssigkeit“ oder das „Wahrnehmungstempo“. In der Regel sind in jedem Untertest des P-S-B 40 Rohwerte maximal zu erreichen. Tabelle 23 gibt einen Auszug aus den altersspezifischen Normen dieses Testes wieder.

Aus der Normentabelle läßt sich der Mittelwert und die Standardabweichung der Rohwerte erschließen. Siebzehnjährige Probanden, die dem P-S-B bei der Eichung unterzogen wurden, erreichten z. B. in Untertest 3, einem Test der „Denkfähigkeit“, im Mittel einen Rohwert von 23. Diesem Rohwert ist dementsprechend der Centilwert 5 zugeordnet; in der Centilskala ($C = 5 + 2 \cdot z$) muß das Mittel bei $C = 5$ liegen. Eine Standardabweichung in dieser Skala („groß“ S) beträgt 2 Centileinheiten.

Lösung zu Übungsaufgabe 9:

$$T = 50 + 10 \cdot \left(\frac{x - \bar{x}}{s} \right) = 50 + 10 \cdot \left(\frac{37 - 39,4}{9,6} \right) \\ = 50 + 10 \cdot (-0,25) = 47,5.$$

Centilwert

17 Jahre

C:	1	2	3	4	5	6	7	8	9
Test									
1+2	16	24	32	40	47	54	61	68	75
3	11	14	17	20	23	26	28	31	33
4	13	16	19	22	24	26	28	30	32
3+4	26	33	38	43	48	52	56	60	64
5	12	18	23	28	32	36	41	45	49
6	11	16	20	23	27	30	33	35	36
5+6	27	37	46	53	60	67	73	78	83
7	0	5	12	20	27	34	38	39	40
8	4	9	14	18	22	25	29	32	35
7+8	10	20	30	40	49	57	64	70	75
9	0	4	9	14	18	22	27	31	37
10	14	23	30	37	44	50	56	59	60
9+10	23	34	44	53	62	72	80	88	95

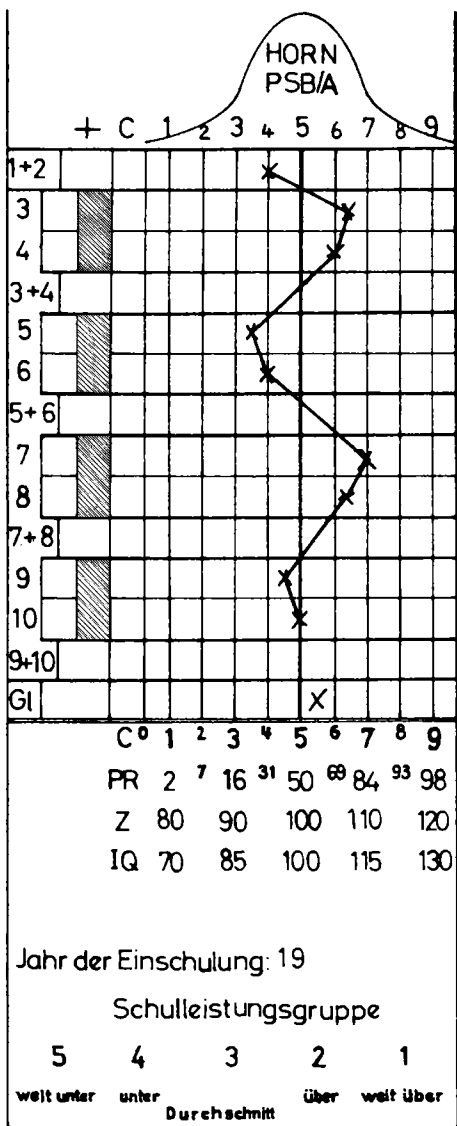
Tab. 23:

Auszug aus der Normtabelle des P-S-B (W. Horn, Prüfungssystem für Schul- und Bildungsberatung P-S-B, Göttingen 1969, S. 25).

Eine Standardabweichung über dem Mittel, bei $C = 7$, finden Sie in Untertest 3 den Rohwert 28. Die Standardabweichung der Rohwerte („klein“ s), die die Siebzehnjährigen in diesem Untertest erreichten, beträgt also $28 - 23 = 5$ Rohwerte.

Beim Untertest 7 beträgt bei 40 maximal erreichbaren Rohwerten der im Mittel erreichte Rohwert 27 ($C = 5$). Hier beträgt die Standardabweichung der Rohwerte $38 - 27 = 11$. (Differenz der Rohwerte zwischen $C = 7$ und $C = 5$.)

Bei der Testauswertung des P-S-B wird über Schablonen die Anzahl der richtig gelösten Aufgaben je Untertest ausgezählt. Zu diesen Rohwerten lassen sich dann für einen Probanden in der entsprechenden Altersnorm die Centilwerte aufsuchen. Die Centilwerte werden anschließend auf dem Testbogen in ein Diagramm eingetragen (Abb. 16). Die verbundenen Centilpunkte ergeben für jeden Probanden ein Leistungsprofil. Das Profil gibt grafisch die Intelligenzstruktur eines Probanden wieder. In Abbildung 16 ist ein fiktives Leistungsprofil zur Verdeutlichung eingetragen.



- Test:
- 1+2 „Allgemeinbildung“
 - 3 „Denkfähigkeit“
 - 4 „Denkfähigkeit“
 - 5 „Worteinfall“
 - 6 „Ratefähigkeit“
 - 7 „Raumvorstellung“
 - 8 „Gliederungsfähigkeit“
 - 9 „Addieren“
 - 10 „Wahrnehmungstempo“

Abb. 16:

Auszug aus dem Testbogen des P-S-B, unter Eintragung eines Leistungsprofils-fiktiv (W. Horn, Prüfungssystem für Schul- und Bildungsberatung, Göttingen 1969).

Neben den altersspezifischen Normen für die Untertests enthält das P-S-B noch eine altersspezifische Norm für die Gesamtleistung (GL) aus allen Untertests. So kann die intellektuelle Leistungsfähigkeit über alle Intelligenzfaktoren hinweg auch noch in einem Gesamtwert ausgedrückt werden.

6.2 Grafische Normen

Neben der Tabellenform lassen sich die Normen eines Testes auch grafisch darstellen. Abbildung 17 gibt die berufsspezifischen Normen der Drahtbiegeprobe von G. A. Lienert wieder¹³. Die Drahtbiegeprobe ist ein standardisierter, manueller Gestaltungstest. Der Test verspricht, das sogenannte Handgeschick zu erfassen. Ein geglühter Eisendraht von 25 cm Länge und 1 mm Durchmesser muß einer Vorlage entsprechend freihändig gebogen werden, ohne den Draht dabei an die Vorlage anzulegen. Die Vorlage der nachzubiegenden Figur ist im Verhältnis von 7:5 gegenüber der Drahtlänge vergrößert. Formrichtigkeit und Sauberkeit sind die wichtigsten Ausführungskriterien. Maximal sind 30 Rohwerte zu erreichen. Von näherem Interesse soll hier nicht die Problematik eines solchen Testes sein, sondern lediglich die Art der Normendarstellung.

Zur Erstellung der berufsspezifischen Normen führten Berufsschüler der verschiedensten Ausbildungsberufe diesen Gestaltungstest durch. Berufsangehörige mit gleicher Leistung wurden in Gruppen zusammengefaßt. In Abbildung 17 ist die Zusammensetzung der Berufsgruppen mit angegeben. Auf der Abszisse in Abbildung 17 sind die Rohwerte abgetragen. Als Transformationsskala ist die Standardwertskala $Z = 100 + 10 \cdot z$ (auch Standardwert (SW) genannt) auf der „rechten“ Ordinate abgetragen. (Anstelle der hier verwendeten Signatur S wird von Lienert σ (Sigma) verwendet.)

Von der Rohwertverteilung jeder Berufsgruppe wurde der Mittelwert $\bar{x}$ und die Standardabweichung s berechnet. Beide Angaben reichen aus, um grafisch die Normen der betreffenden Berufsgruppe darzustellen: Dem Mittelwert $\bar{x}$ der Rohwerte einer Berufsgruppe wird das Mittel der Skala $Z = 100$ zugeordnet. Bei dem Rohwert, der eine Standardabweichung s unter dem Mittelwert $\bar{x}$ liegt, muß der Z-Wert 90 liegen. Der Z-Wert 90 liegt nämlich in der Z-Skala gleichfalls eine Standardabweichung unter dem Mittel. Diese zwei Punkte im Liniennetz reichen aus, um eine Gerade zu zeichnen.

Die Berufsschüler der Gruppe 5 (Feinmechaniker u. a.) erreichen die höchsten Leistungen. Das Mittel $Z = 100$ entspricht einem Rohwert von ca. 21. Das heißt, daß der Mittelwert der Rohwerte dieser Gruppe $\bar{x} = 21$ beträgt.

Zusammensetzung der Berufsgruppen

Die einzelnen Berufsgruppen umfassen folgende Berufe:

Gruppe 1: Maler, Landwirte, Hilfsarbeiter, Postjungboten, Bäcker, N = 111.

Gruppe 2: Kraftfahrzeugmechaniker, Dreher, Heizungsmonteure, Dachdecker, Maurer, Weber und Schneider, kaufmännische Lehrlinge, N = 165.

Gruppe 3: Maschinenschlosser, Modellbauer und -schreiner, Schreiner, N = 255.

Gruppe 4: Werkzeugmacher, Fliesenleger, Polsterer und Tapezierer, Gastwirte und Köche etc., Verwaltungsangestellte, N = 93.

Gruppe 5: Feinmechaniker, Elektromechaniker, Fahrzeugbauer, Ofensetzer, N = 114.

Bei den Gruppeneinteilungen wurde darauf geachtet, daß die einzelnen Stichproben hinreichend groß und die Mittelwerte der zusammengefaßten Berufe nicht signifikant verschieden waren. Das gleiche galt auch für die Zusammenfassung von Schüler- und Altersgruppen.

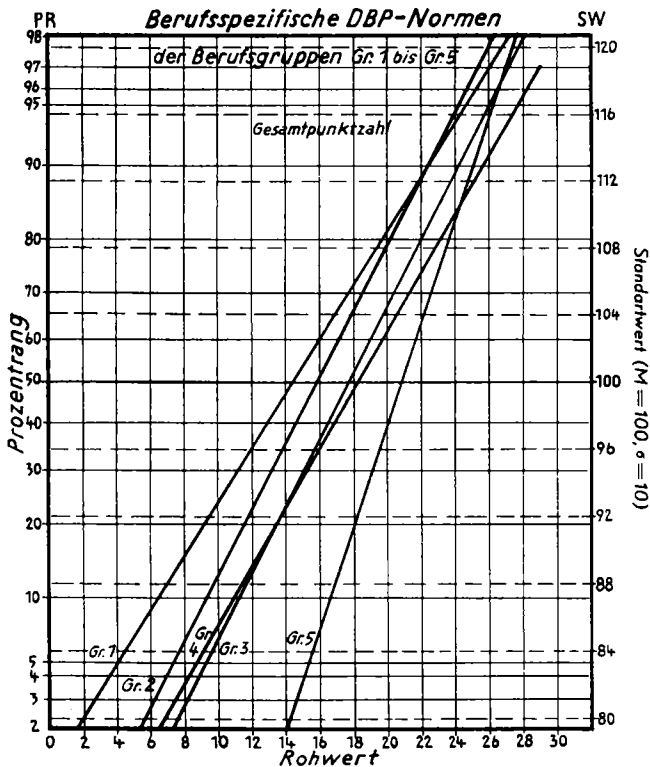


Abb. 17:

Auszug aus den Normen der Drahtbiegeprobe (D-B-P). (G. A. Lienert, Die Drahtbiegeprobe als standardisierter Test: Handanweisung für die Durchführung, Interpretation und Auswertung, Göttingen 1967, S. 26 u. 21.)

Im Vergleich dazu findet sich für die Gruppe 1 (Maler u. a.) bei $Z = 100$ ein Rohwert von ca. 14. Das heißt, daß die Berufsschüler dieser Gruppe im Schnitt 14 Rohwerte erreichten. Sie schneiden in diesem Test deutlich schlechter ab als die Feinmechaniker u. a. der Gruppe 5.

In einer grafisch angelegten Norm läßt sich wie aus einer Normentabelle zu jedem Rohwert ein Standardwert ablesen. Erreicht z.B. ein Berufsschüler im Ausbildungsberuf Feinmechaniker in der Drahtbiegeprobe einen Rohwert von 16, entspricht dies einem Z -Wert von 86 in seiner Vergleichsgruppe. Gemessen an seiner Vergleichsgruppe ist die Leistung dieses Probanden ausgesprochen gering. Sie liegt mehr als eine Standardabweichung unter dem Mittelwert. Der Z -Wert 86 entspricht einem Prozentrang von ca. 8. 92 % der Schüler der Vergleichs- oder Eichgruppe erzielten bessere Leistungen als er. Vergleichen Sie dagegen die Leistung dieses Probanden mit den Leistungen, welche die Angehörigen der Gruppe 4 (Werkzeugmacher u. a.) erreichten, stellt sich das Bild anders dar. In dieser Gruppe entspricht einem Rohwert von 16 ein Z -Wert von ca. 96. Der korrespondierende Prozentrang beträgt $PR = 35$. 65 % in dieser Vergleichsgruppe erreichen bessere Leistungen (gegenüber 92 % bei der Vergleichsgruppe Feinmechaniker u. a.).

Die Drahtbiegeprobe enthält neben den hier vorgestellten berufsspezifischen Normen noch Alters- und schulspezifische Normen.

6.3 Normierung (Standardisierung) eines informellen Testes

In der gleichen Weise wie erwerbbare Tests normiert werden, läßt sich auch der vom Lehrer selbst erstellte Schulleistungstest normieren (standardisieren, eichen).

Eingangs wurde der informelle Leistungstest Vitamine und Mineralstoffe vorgestellt. Hierbei handelte es sich um die erste Testfassung. Der Test wird nun revidiert, indem mehrdeutige Aufgaben umgeschrieben werden. Neue Aufgaben werden zur Ausschöpfung der Leistungsfähigkeit der Schüler eingefügt (siehe unten die Aufgabenanalyse). Danach wird der Unterricht über Vitamine und Mineralstoffe in ähnlichen Klassen wiederholt, sei es sofort oder im darauf folgenden Unterrichtsjahr. Der Test wird dabei mehrmals neu eingesetzt und revidiert. Nach dieser Entwicklungsphase wird eine Testfassung, die der Lehrer als zufriedenstellend und endgültig ansieht, zur Eichung in mehreren Klassen eingesetzt. Insgesamt mögen dabei 100 Schüler getestet werden. Von den Rohwerten dieser 100 Schüler wird der Mittelwert $\bar{x}$ und die Standardabweichung s berechnet. Verteilen sich die Rohwerte annähernd normal (vgl. die qualitative Prüfung der Normalität S. 56), läßt sich die Normen-

tabelle anlegen. Bei Schulleistungstests empfiehlt es sich, die T-Skala zu benutzen:

$$T = 50 + 10 \cdot z = 50 + 10 \left[\frac{x - \bar{x}}{s} \right].$$

Der Mittelwert der Rohwerte möge dabei $\bar{x} = 12$, die Standardabweichung $s = 3$ betragen. Insgesamt möge die endgültige Testfassung des Testes Vitamine und Mineralstoffe 20 Aufgaben umfassen. Je Aufgabe wird ein Rohwert vergeben, so daß maximal 20 Rohwerte zu erreichen sind. Zu jedem der 20 möglichen Rohwerte (RW) wird nun der entsprechende T-Wert berechnet. In einer Tabelle lassen sich den Rohwerten von 0–20 die entsprechenden T-Werte zuordnen. Für $RW = 18$ beträgt dann z. B. der T-Wert:

$$T = 50 + 10 \cdot \frac{(18 - 12)}{3} = 70.$$

Wird nun später in vergleichbaren Klassen wieder der Test Vitamine und Mineralstoffe eingesetzt, so läßt sich die Leistung jedes einzelnen Schülers mit der Leistung einer größeren Vergleichsgruppe vergleichen. Mit anderen Worten, zu dem ermittelten Rohwert eines einzelnen Schülers wird in der selbst angelegten Normentabelle des Testes der entsprechende T-Wert aufgesucht, der die Leistung des Schülers im Vergleich zur Norm von 100 Schülern ausdrückt.

7. Korrelation

7.1 Das Konzept der Korrelation

Eine Korrelation drückt die „co-relation“, d. h. die zusammengehende Beziehung zwischen zwei Merkmalen aus. Bei männlichen Personen besteht z. B. ein Zusammenhang zwischen der Körpergröße und dem Körpergewicht. Größere Personen wiegen in der Regel mehr als kleinere. Allerdings ist kaum zu erwarten, daß die größte Person die schwerste, die zweitgrößte die zweitschwerste, die drittgrößte die drittschwerste ist, usw., bis zur kleinsten Person. Ein Zusammenhang zwischen den beiden Merkmalen Größe und Gewicht besteht nur zu einem bestimmten Ausmaß. Die Frage ist, wie stark dieser Zusammenhang ist, m. a. W. wie stark die zusammengehende Beziehung ist.

Um diese Frage zu beantworten, muß für eine Gruppe von Personen an jeder Person die Körpergröße und das Körpergewicht, also beide Merkmalsgrößen bestimmt werden (Tabelle 24).

Person	x Körpergröße (cm)	y Körpergewicht (kg)
A	180	78
B	175	73
C	178	78
D	170	69
E	182	86
.	.	.
.	.	.
.	.	.

Tab. 24:
Körpergröße und Körpergewicht bei männlichen Personen.

Die Körpergröße kann als die Variable x und das Körpergewicht als die Variable y bezeichnet werden. Die Variablen x und y korrelieren (positiv), wenn hohe Werte in der Variablen x dazu tendieren, mit hohen Werten in der Variablen y einherzugehen. Der Korrelationskoeffizient r ist nun ein rechnerisches Maß für den Zusammenhang zweier Variablen. Er drückt numerisch den Grad der zusammengehenden Beziehung zwischen zwei Merkmalen aus. Der Koeffizient kann Werte zwischen $r = +1,00$ und $r = -1,00$ annehmen. Sein Vorzeichen drückt die Richtung der Beziehung aus.

Ein Koeffizient von $r_{xy} = +1,00$ bedeutet, daß ein vollständiger und gleichsinniger Zusammenhang zwischen den Variablen x und y besteht. In Tabelle 24 würde dies bedeuten, daß immer die größere Person auch das größere Gewicht hat. Man sagt, daß $r = +1,00$ eine perfekt positive Be-

ziehung zwischen zwei Variablen darstellt. Anstelle von positiver Beziehung wird auch von direkter Beziehung gesprochen.

Ein Korrelationskoeffizient von $r_{xy} = -1,00$ besagt, daß ein vollständiger, aber gegenläufiger Zusammenhang zwischen den Variablen x und y besteht. In Tabelle 24 würde dies bedeuten, daß immer die größere Person auch das kleinere Gewicht hat. Die größte Person würde dann das geringste Gewicht aufweisen. $r = -1,00$ stellt eine perfekt negative; man sagt auch eine perfekt inverse Beziehung zwischen zwei Variablen dar.

Ein Koeffizient von $r_{xy} = 0,00$ sagt aus, daß kein systematischer Zusammenhang zwischen den Variablen x und y besteht. In Tabelle 24 würde dies bedeuten, daß große Personen von sowohl großem als auch von kleinem Gewicht sind. Ebenso würden die kleineren Personen in Relation zu ihrer Körpergröße einmal weniger (direkte Beziehung) aber auch einmal mehr (inverse Beziehung) wiegen. Besteht zwischen zwei Variablen zu gleichen Anteilen sowohl eine direkte als auch eine inverse Beziehung, ist der Korrelationskoeffizient $r = 0,00$. Eine systematische Beziehung besteht zwischen beiden Variablen nicht.

Definition 17:

Die *Korrelation* drückt den Zusammenhang zwischen zwei Variablen aus. Der Grad des Zusammenhanges wird durch den Korrelationskoeffizienten r angegeben. Sein Vorzeichen drückt die Richtung des Zusammenhanges aus.

Es gilt: $-1,00 \leq r \leq +1,00$

Dabei steht

$r = +1,00$ für eine perfekt direkte (positive) Beziehung zwischen zwei Variablen,

$r = 0,00$ für keine systematische Beziehung zwischen zwei Variablen und

$r = -1,00$ für eine perfekt inverse (negative) Beziehung zwischen zwei Variablen.

Grafische Veranschaulichung des Korrelationskonzepts

In Abbildung 18 sind verschiedene korrelative Beziehungen zwischen zwei Variablen über Streuungsdiagramme verdeutlicht.

Abbildung 19 gibt einen Fall wieder, in der eine Beziehung zwischen den Variablen x und y besteht; allerdings kann keine Gerade gezogen werden, die diese Beziehung annähernd wiedergeben kann. Die Beziehung zwischen x und y ist nicht gradlinig sondern kurvilinear. Für Streuungsdiagramme dieser Art sind die in Kap. 7.2 vorgestellten Berechnungsmethoden für r nicht angemessen.

Bei einer korrelativen Beziehung von $r = +1,00$ zwischen zwei Leistungsvariablen ist die Rangreihe der Personen in der Variablen x identisch mit der Rangreihe in der Variablen y. Die Person, die in der Variablen x die höchste Leistung erreicht, erzielt auch in der Variablen y die höchste Leistung. Sie erhält in der Variablen x und in der Variablen y den Rang 1. Jene Person, die in der Variablen x die zweitbeste Leistung erbringt, dort den Rang 2 erhält, muß bei einem $r = +1,00$ auch in der Variablen y die zweitbeste Leistung erreichen, Rang 2 (Abbildung 20a).

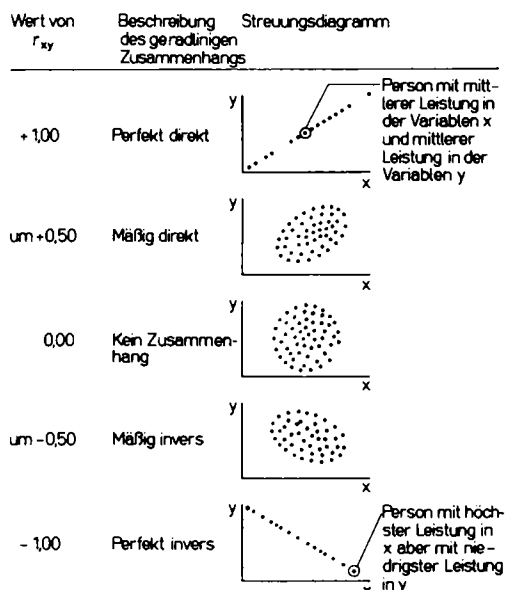


Abb. 18:

Veranschaulichung verschiedener korrelativer Beziehungen über Streuungsdiagramme (nach G. V. Glass, J. C. Stanley, Statistical Methods in Education and Psychology, Englewood Cliffs, New Jersey, 1970, S. 117).

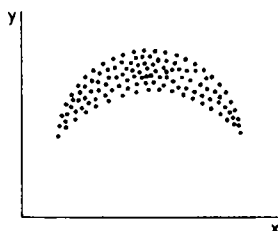


Abb. 19:

Veranschaulichung einer kurvilinearen Beziehung zwischen zwei Variablen (nach J. C. Stanley, K. D. Hopkins, Educational and Psychological Measurement and Evaluation, Englewood Cliffs, New Jersey 1972, S. 61).

Nimmt die Person, die in der Variablen x auf dem Rang 1 steht (höchste Leistung), in der Variablen y den letzten Rang ein (geringste Leistung) und setzt sich diese Gesetzmäßigkeit fort, liegt die perfekt inverse Beziehung von $r = -1,00$ vor (Abbildung 20c). Für ein $r = 0,00$ ergibt sich in Abbildung 20b ein unregelmäßiges Schnittfeld der Verbindungslinien.

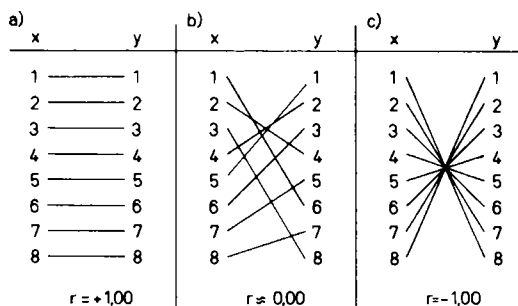


Abb. 20:

Veranschaulichung korrelativer Beziehungen über Rangreihen (nach P. Birkel, Glossar wichtiger testtheoretischer Begriffe, in: Test Informationen: Hinweise zur Auswahl von Tests, Beltz Test Gesellschaft, Weinheim 1976, S. 40, modifiziert).

7.2 Berechnungsverfahren des Korrelationskoeffizienten

Im folgenden werden Rechenverfahren vorgestellt, die aus dem Pearson-Produkt-Moment-Korrelationskoeffizienten abgeleitet sind¹⁴. (Dieser Korrelationskoeffizient wird auch unter der Bezeichnung „Pearson-Bravaischer Maßkorrelationskoeffizient“ geführt.)

Ein einfaches und formelmäßig leicht zu behaltendes Rechenverfahren zur Bestimmung des Korrelationskoeffizienten verfährt über die z-Werte der zu korrelierenden Variablen.

Definition 18:

Der *Korrelationskoeffizient* ist gleich dem Durchschnittsprodukt der z-Werte zweier Variablen x und y.

$$r = \frac{\sum z_x \cdot z_y}{N}$$

mit z_x , z_y z-Werte in den Variablen x und y

N Zahl der Rohwertträger in der Variablen x oder y

In Tabelle 25 ist nach dieser Formel eine Berechnung des Korrelationskoeffizienten durchgeführt worden. 25 Schüler haben sich nach einer Unterrichtsstrecke einem Leistungstest I (Variable x) und nach einer weiteren Unterrichtsstrecke einem Leistungstest II (Variable y) unterzogen. In beiden Tests sind maximal 50 Rohwerte zu erreichen. Schüler A erreichte im Test I (x) 35 und im Test II (y) 38 Rohwerte. Mit angegeben sind die Standardabweichung und der Mittelwert der Rohwertsätze x und y. In Tabelle 25b sind die Rohwerte aus Tabelle 25a in z-Werte transformiert worden. Schüler A erhält einen z_x -Wert von:

$$z_{Ax} = \frac{35 - \bar{x}}{s_x} = \frac{35 - 34,16}{7,97} = 0,11.$$

Sein z-Wert in y beträgt:

$$z_{Ay} = \frac{38 - \bar{y}}{s_y} = \frac{38 - 34,28}{7,27} = 0,51.$$

Wäre die Korrelation zwischen den Testleistungen perfekt direkt ($r_{xy} = +1,00$), würde der z_x - mit dem z_y Wert bei jedem Schüler identisch sein, d. h. es würde jeder Schüler in beiden Tests eine gleich große Leistung er-

a
b

Schüler	Test I Rohwert (x)	Test II Rohwert (y)	Test I z _x	Test II z _y
A	35	38	0,11	0,51
B	38	45	0,48	1,47
C	22	25	-1,53	-1,28
D	40	33	0,73	-0,18
E	45	44	1,36	1,34
F	31	25	-0,40	-1,28
G	27	40	-0,90	0,79
H	42	40	0,98	0,79
I	40	32	0,73	-0,31
J	30	22	-0,52	-1,69
K	28	29	-0,77	-0,73
L	41	38	0,86	0,51
M	29	32	-0,65	-0,31
N	35	40	0,11	0,79
O	46	48	1,49	1,89
P	22	28	-1,53	-0,86
Q	21	30	-1,65	-0,59
R	33	26	-0,15	-1,14
S	37	39	0,36	0,65
T	41	42	0,86	1,06
U	26	30	-1,02	-0,59
V	22	31	-1,53	-0,45
W	33	22	-0,15	-1,69
X	42	40	0,98	0,79
Y	48	38	1,74	0,51

$$N = 25$$

$$\bar{x} = 34,16 \quad \bar{y} = 34,28 \quad s_x = 7,97 \quad s_y = 7,27$$

$$r = \frac{\sum z_x \cdot z_y}{N} = \frac{16,5553}{25} = 0,66$$

Tab. 25:

Berechnung des Korrelationskoeffizienten r über z-Werte.

bringen. Ein kurzer Blick auf Tabelle 25 zeigt bereits, daß dies nicht der Fall ist. Die Leistung variiert bei manchen Schülern erheblich. Die Werte scheinen nicht gerade hoch miteinander zu „covariieren“. Eine große Änderung liegt beim Schüler G vor: von $z_x = -0,90$ in der Variablen x bis $z_y = +0,79$ in der Variablen y. Eine geringe Leistungsänderung weist Schüler E auf: von $z_x = +1,36$ nach $z_y = +1,34$.

Zur Berechnung des Korrelationskoeffizienten wird nun in jeder Zeile der Tabelle 25b der z_x - mit dem z_y Wert multipliziert. Dabei ist auf das Vorzeichen zu achten ($(-)\cdot(-) = (+)$; $(-)\cdot(+) = (-)$; $(+)\cdot(-) = (-)$ und $(+)\cdot(+) = (+)$). Werden diese Produkte aufsummiert ($\sum z_x \cdot z_y$) und durch $N = 25$ dividiert, so ist der Korrelationskoeffizient:

$$r = \frac{\sum z_x \cdot z_y}{N} = \frac{15,8258}{25} = \underline{\underline{0,66}}$$

Der Korrelationskoeffizient läßt sich auch direkt aus den Rohwerten zweier Variablensätze berechnen. Die Formel für r, wie sie in Definition 18 angegeben ist, wird dann jedoch komplizierter. Die „Maschinenformel“ zur Berechnung von r, die sich für den Einsatz eines elektronischen Taschen- oder Tischrechners, eignet, lautet:

$$r = \frac{N \cdot (\sum x \cdot y) - (\sum x) \cdot (\sum y)}{\sqrt{[N \cdot \sum x^2 - (\sum x)^2] \cdot [N \cdot \sum y^2 - (\sum y)^2]}}$$

In Tabelle 26a und b ist für zwei einfache Rohwertsätze x und y die Korrelation nach dieser Formel berechnet worden. Die Funktionszeichen des Tischrechners (Walther ETR 25), wie sie im Ausdruck in Tabelle 26 erscheinen, sind am Fuß der Tabelle erklärt (siehe hierzu auch S. 26). Modernere Taschen- oder Tischrechner haben die Korrelationsberechnung r als Funktion fest einprogrammiert, so daß nur die Rohwerte beider Variablen eingetippt werden müssen und durch entsprechenden Tastendruck die Korrelationsberechnung automatisch abläuft. In Tabelle 26c ist

Person	Variable x	Variable y
1	20	8
2	10	12
3	8	6
4	3	5
5	16	21
6	18	14
7	21	10
8	12	16
9	9	13
10	8	8

Tab. 26a:

Rohwertsatz zur Berechnung des Korrelationskoeffizienten r nach der „Maschinenformel“.

dies für das gleiche Datenmaterial demonstriert. Verwandt wurde wieder der elektronische Taschenrechner der Marke Texas Instruments TI – 51 – III.

20.00 ×	8.00 +	8.00 ×
8.00 =	12.00 +	8.00 =
160.00 Σ	6.00 +	64.00 Σ
10.00 ×	5.00 +	12.00 ×
12.00 =	21.00 +	12.00 =
120.00 Σ	14.00 +	144.00 Σ
8.00 ×	10.00 +	6.00 ×
6.00 =	16.00 +	6.00 =
48.00 Σ	13.00 +	36.00 Σ
3.00 ×	8.00 =	5.00 ×
5.00 =	Σ y = 113.00 *	5.00 =
15.00 Σ		25.00 Σ
16.00 ×		21.00 ×
21.00 =	20.00 ×	21.00 =
336.00 Σ	20.00 =	441.00 Σ
18.00 ×	400.00 Σ	14.00 ×
14.00 =	10.00 ×	14.00 =
252.00 Σ	10.00 =	196.00 Σ
21.00 ×	100.00 Σ	10.00 ×
10.00 =	8.00 ×	10.00 =
210.00 Σ	8.00 =	100.00 Σ
12.00 ×	64.00 Σ	16.00 ×
16.00 =	3.00 ×	16.00 =
192.00 Σ	3.00 =	256.00 Σ
9.00 ×	9.00 Σ	13.00 ×
13.00 =	16.00 ×	13.00 =
117.00 Σ	16.00 =	169.00 Σ
8.00 ×	256.00 Σ	8.00 ×
8.00 =	18.00 ×	8.00 =
64.00 Σ	18.00 =	64.00 Σ
Σ x y = 1514.00 *	324.00 Σ	Σ y ² = 1495.00 *
	21.00 ×	
	21.00 =	
20.00 +	441.00 Σ	
10.00 +	12.00 ×	
8.00 +	12.00 =	
3.00 +	144.00 Σ	
16.00 +	9.00 ×	
18.00 +	9.00 =	
21.00 +	81.00 Σ	
12.00 +	8.00 ×	
9.00 +	8.00 =	
8.00 =	64.00 Σ	
Σ x = 125.00 *	Σ x ² = 1883.00 *	

$$\begin{aligned} & 125.00 \times \\ & 113.00 = \\ (\Sigma x) \cdot (\Sigma y) &= 14125.00 \end{aligned}$$

$$\begin{aligned} & 10.00 \times \\ & 1514.00 = \\ N \cdot (\Sigma xy) &= 15140.00 \end{aligned}$$

$$\begin{aligned} & 15140.00 - \\ & 14125.00 = \\ a &= 1015.00 \end{aligned}$$

$$\begin{aligned} & 10.00 \times \\ & 1883.00 = \\ N \Sigma x^2 &= 18830.00 \end{aligned}$$

$$\begin{aligned} & 125.00 \times \\ & 125.00 = \\ (\Sigma x)^2 &= 15625.00 \end{aligned}$$

$$\begin{aligned} & 18830.00 - \\ & 15625.00 = \\ b &= 3205.00 \end{aligned}$$

$$\begin{aligned} & 3205.00 \times \\ & 2181.00 = \\ b \cdot c &= 6990105.00 \\ & 6990105.00 \sqrt{} = \\ \sqrt{b \cdot c} &= 2643.88 \end{aligned}$$

$$\begin{aligned} & 1015.00 \div \\ & 2643.88 = \\ r &= 0.38 \end{aligned}$$

Zeichenerklärung:

= "gleich", × "Mal", · "Ergebnis", √ "Wurzel", ∓ "Speicher",
÷ "geteilt durch"

$$r = \frac{N \cdot (\Sigma xy) - (\Sigma x) \cdot (\Sigma y)}{\sqrt{[N \cdot \Sigma x^2 - (\Sigma x)^2] \cdot [N \cdot \Sigma y^2 - (\Sigma y)^2]}}$$

$$\begin{aligned} & 10.00 \times \\ & 1495.00 = \\ N \Sigma y^2 &= 14950.00 \end{aligned}$$

$$\begin{aligned} & 113.00 \times \\ & 113.00 = \\ (\Sigma y)^2 &= 12769.00 \end{aligned}$$

$$\begin{aligned} & 14950.00 - \\ & 12769.00 = \\ c &= 2181.00 \end{aligned}$$

Tab. 26b:
Berechnung des Korrelationskoeffizienten r nach der Maschinenformel (Daten aus
Tabelle 26a).

Tasten	Anzeige/Bemerkungen	Tasten	Anzeige/Bemerkungen
2nd → CA	0 Gesamtlöschung	18 → x↔y	
20 → x↔y	0. Eingabe von x ₁	14 → Σ+	
8 Σ+	1. Eingabe von y ₁	21 → x↔y	
10 → x↔y		10 → Σ+	
12 → Σ+		12 → x↔y	
8 → x↔y		16 → Σ+	
6 → Σ+		9 → x↔y	
3 → x↔y		13 → Σ+	
5 → Σ+		8 → x↔y	
16 → x↔y		8 → Σ+	10. Eingabe von y ₁₀
21 → Σ+		2nd → Corr	0,3839054 (Korrelationskoeffizient)

Tab. 26c:

Berechnung des Korrelationskoeffizienten r über einen elektronischen Taschenrechner (Texas Instruments TI-51-III) mit fest einprogrammierter Korrelationsfunktionstaste (Daten aus Tabelle 26a).

Tastenerklärung:

2nd : (second) Zugang zur Tastaturebene 2

CA : Clear all

x↔y : x-y Austausch Taste, Eingabe der x-Werte für die lineare Regression

Σ+ : Sigma-Taste, Eingabe der y-Werte für die lineare Regression

2nd Corr : Korrelation

7.3 Interpretation des Korrelationskoeffizienten

Über folgende Beziehung läßt sich der Korrelationskoeffizient anschaulich verdeutlichen. Es kann gezeigt werden, daß

$$z'_y = r \cdot z_x \text{ ist,}$$

mit

z'_y erwarteter oder vorherzusagender z-Wert der Variablen y („Kriteriumsvariable y“),

r Korrelationskoeffizient zwischen Variablen x und y und,

z_x bekannter z-Wert der Variablen x („Prediktorvariable x“).

Zwischen der Leistung im Hochsprung (y) und der im Weitsprung (x) möge für eine Gruppe von Kindern eine korrelative Beziehung von

$r = 0,60$ bestehen. Dies sei über eine Meßreihe festgestellt worden. Die Leistung im Hochsprung läßt sich nun aus der Leistung im Weitsprung vorhersagen. Erreicht ein Kind A eine Weitsprungleistung von $z_x = + 2$, so wird seine geschätzte Hochsprungleistung

$$z'_y = r \cdot z_x = 0,6 \cdot 2 = 1,2$$

betragen.

Während die Weitsprungleistung des Kindes A zwei z-Werteinheiten, d. h. 2 Standardabweichungen über dem Mittelwert der Vergleichsgruppe liegt, wird seine Hochsprungleistung (nach der Vorhersage) nur das 1,2fache einer Standardabweichung über dem Mittelwert der Vergleichsgruppe liegen. Dies sind 60% des Abstandes, den Person A im Weitsprung vom Mittelwert entfernt liegt. Es läßt sich somit für die Interpretation des Korrelationskoeffizienten sagen:

Definition 18a:

Der *Korrelationskoeffizient* r gibt an, zu wieviel Prozent die Leistung in der vorherzusagenden Variablen y so weit von ihrem Mittel $\bar{y}$ entfernt ist, wie die Leistung in der bekannten Variablen x von ihrem Mittel $\bar{x}$ entfernt ist.

Wenn die korrelative Beziehung zwischen der Weit- (x) und Hochsprungleistung (y) $r = 0,6$ beträgt, dann wird eine Leistung im Hochsprung zu 60% so weit von der mittleren Hochsprungleistung entfernt sein, wie die Leistung im Weitsprung von der mittleren Weitsprungleistung entfernt ist. Betrüge hier $r = 1,00$, so würde eine Leistung im Hochsprung zu 100% so weit von der mittleren Hochsprungleistung entfernt sein, wie eine Leistung im Weitsprung von der mittleren Weitsprungleistung entfernt ist.

Ein Korrelationskoeffizient kann also nicht als ein Prozentwert im üblichen Sinne aufgefaßt werden. Ein r von 0,66 gibt nicht eine Übereinstimmung von 66 Prozent wieder. Soll von Prozentsätzen der Übereinstimmung bei der Interpretation einer Korrelation gesprochen werden, kann dies nur im oben angegebenen Sinne geschehen.

In der Regel wird eine Korrelation von $r \geq 0,8$ als hoch, eine Korrelation von $r \geq 0,6$ als mäßig hoch und eine von $r \leq 0,5$ als mittel bis niedrig angesehen. Solche Angaben sind jedoch vage. Eine Korrelation von $r = 0,6$ kann einmal als hoch und einmal als niedrig gelten. Eine Korrelation von $r = 0,6$ zwischen der Intelligenz, gemessen an einem Intelligenztest, und der Schulleistung, gemessen an den Schulnoten, ist hoch. Gewöhnlich fallen solche Korrelationen niedriger aus. Eine Korrelation von $r = 0,6$ ist dagegen niedrig für eine Wiederholungsmessung mit ein und demselben

Intelligenztest bei gleichen Probanden. – Die Interpretation der Höhe eines Korrelationskoeffizienten ist relativ. Es kommt darauf an, wie Korrelationen bei ähnlichem Datenmaterial im allgemeinen ausfallen. Man muß bereits eine Erwartung über die Richtung und die absolute Größe der Korrelation haben.

Für die Interpretation einer Korrelation ist ferner bedeutsam, wie groß das N , die Zahl der Wertepaarträger ist, für die der Korrelationskoeffizient berechnet worden ist. Nehmen Sie z. B. an, es wird das Körpergewicht und die Körpergröße bei $N = 2$ Personen miteinander korreliert. Der berechnete Koeffizient r wird entweder $+1,0$ oder $-1,0$ betragen (es sei denn beide Personen hätten exakt die gleiche Körpergröße oder haben beide das gleiche Körpergewicht). Die im Körpergewicht höher liegende Person wird entweder die größere der beiden (dann $r = +1,0$) oder die kleinere von beiden (dann $r = -1,0$) sein. Bei geringem N kann ein Korrelationskoeffizient irreführend sein. Das N muß in der Regel deutlich größer als 10 sein, um noch sinnvoll eine Korrelation berechnen zu können.

7.4 Korrelation und Kausalität

Wenn zwischen zwei Variablen eine Korrelation besteht, dann muß nicht notwendigerweise eine kausale Beziehung zwischen den beiden Variablen bestehen, z. B. daß x die Ursache für y ist. An einem einfachen Beispiel mag dies deutlich werden: Es besteht vermutlich eine Korrelation zwischen der Zahl der tief fliegenden Schwalben und den Millimetern an Regen in den nächsten drei Stunden. Aber dies bedeutet nicht, daß die Schwalben den Regen verursachen.

Wenn tatsächlich die eine Variable die andere verursacht, dann sind beide meist stark miteinander korreliert. Aber das Umgekehrte muß nicht wahr sein. Wenn zwei Variablen miteinander korrelieren, dann kann, muß aber nicht, die eine die andere Variable verursachen.

Wenn zwei Variablen hoch miteinander korreliert sind und Sie daraus ableiten, daß die eine die andere verursacht, so können Sie einen falschen Schluß gezogen haben. Auf der Basis der Korrelation allein, ist dieser Schluß nicht zu ziehen. Nehmen Sie z. B. an, bei Schülern betrage die Korrelation zwischen der Variablen „Angst“ bei der Durchführung eines Mathematiktestes (x) und der Leistung in diesem Mathematiktest (y) $r = -0,70$. Dies würde bedeuten, daß Schüler mit großer Angst von geringer Mathematikleistung und Schüler mit geringer Angst von hoher Mathematikleistung sind. Hat also die Variable x (Angst) die Variable y (Mathematikleistung) verursacht? Oder ist die Mathematikleistung der Faktor, der die Angst verursacht? Schüler mit geringer Leistung in Mathematik könn-

ten ängstlich werden, wenn ihre Leistung in Mathematik getestet wird. In diesem Falle würde die Variable y (Mathematikleistung) die Variable x (Angst) hervorgerufen haben. (Hier könnte ein Experiment weiterhelfen, bei dem eine Gruppe in „Angst versetzt“ wird und deren Leistung in Mathematik mit der einer Gruppe ohne Angst verglichen wird, sofern ein solches Experiment moralisch vertretbar ist.)

Formal gesehen lassen sich vier Möglichkeiten des Variablenzusammenhangs nennen, wenn zwei Variablen x und y korrelieren (Abbildung 21).

Zu a:

Die Werte in y können durch die Werte von x bestimmt sein.

Zu b:

Die Situation kann auch umgekehrt sein, indem y die Werte in x verursacht.

Zu c:

Denkbar ist auch, daß ein dritter Faktor existiert, der auf x und y einwirkt und so die zusammengehende Beziehung zwischen beiden hervorruft.

Zu d:

Noch komplizierter wird es, wenn der „dritte Faktor“ bei Fall c nicht ein einziges Charakteristikum darstellt, sondern ein ganzes Gefüge von Variablen, die miteinander in Beziehung stehen.

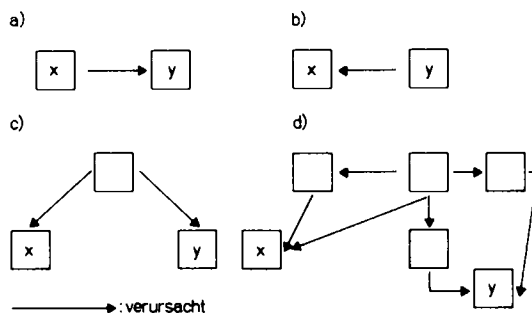


Abb. 21:

Die möglichen kausalen Beziehungen zwischen korrelierenden Variablen x und y (nach J. M. Diehl, H. U. Kohr, Statistische Methoden: Band I Deskriptive Statistik, 3. verb. Aufl., Fachbereich Psychologie, Justus Liebig-Universität Gießen, o.J., S. 98, modifiziert).

Mit diesen formalen Überlegungen seien hier die Betrachtungen zur Problematik „Korrelation und Kausalität“ abgeschlossen. Zusammenfassend sei noch einmal hervorgehoben: Das Bestehen einer Korrelation zwischen zwei Variablen *kann* bedeuten, daß eine monokausale Beziehung zwischen den beiden Variablen besteht, in welcher Richtung auch immer. Aus dem Bestehen der Korrelation allein jedoch kann nicht auf die Kausalität zwingend gefolgert werden.

Teil B

8. Validität

8.1 Der Begriff der Validität (Gültigkeit)

Ein Test ist valide¹⁵, wenn er tatsächlich das mißt, was er zu messen vorgibt. Wenn z.B. in einem Rechentest Textaufgaben so formuliert sind, daß sie hohe Anforderungen an das Leseverständnis stellen, wird nicht nur die Fähigkeit im angewandten Rechnen, sondern auch das Leseverständnis geprüft. Der Test würde neben dem, was er eigentlich messen soll, noch etwas anderes messen.

Die Validität gibt an, wie gut der Test die Funktion erfüllt, für die er eingesetzt wird. Bei der Validität lassen sich drei verschiedene Aspekte unterscheiden, auf die im folgenden kurz eingegangen werden soll: die Inhaltsvalidität, die kriterienbezogene Validität und die Konstruktvalidität.

8.2 Inhaltsvalidität

Bei Schulleistungstests, seien es informelle, vom Lehrer selbst erstellte oder erwerbbar Tests, ist die Inhaltsvalidität der relevante Typ der Validitätsbestimmung.

Bei diesen Tests ist zu prüfen, ob der Test auch das mißt, was gelehrt wurde:

- Gibt der Test die Lernziele des Unterrichts wieder? – Die Lernziele müssen nach ihrem Inhalt und ihrem zugrundeliegenden Verhalten in den Testaufgaben Entsprechungen finden.
- Besteht ein Gleichgewicht zwischen den Schwerpunkten in der Lernzielvorgabe und den Schwerpunkten im Test? – Wenn z.B. bestimmte Lernziele ausführlich im Unterricht verfolgt wurden, müssen im Test auch diese Lernziele mit mehr Aufgaben abgedeckt werden.
- Setzt der Test Fähigkeiten voraus, die mit dem, was geprüft werden soll, in keinem Zusammenhang stehen? – Es müssen z.B. Leseverständlichkeit des Testes und der Wortschatz, auf den er abstellt, für die Testbearbeiter angemessen sein.

Bei erwerbbar Schulleistungstests haben Experten beurteilt, ob die Testaufgaben die Lernziele erfassen, die in den Lehrplänen für die entsprechende Altersstufe gefordert werden. Stimmen die Experten überein, wird auch von der Lehrplanvalidität eines Testes gesprochen.

Die Lehrplanvalidität bedeutet jedoch nicht notwendigerweise, daß ein Test für jeden Einsatz inhaltsvalide ist. Die Inhaltsvalidität muß jeweils neu überprüft werden, indem eine Sichtung der Testaufgaben vorge-

nommen wird. Auf die Frage „Ist der Test inhaltsvalide?“ gibt es in der Regel keine direkte Antwort. Es kommt auf den Zweck an, für den er eingesetzt werden soll.

8.3 Kriterienbezogene Validität

Die kriterienbezogene Validitätsbestimmung verfährt empirisch und wird aus diesem Grunde auch als empirische Validität bezeichnet.

Ein Validitätskoeffizient wird über folgendes Verfahren bestimmt: Probanden bearbeiten den Test, der zu validieren ist. Darüber hinaus werden sie noch einem anderen Test unterzogen, der bereits valide jenes Merkmal mißt, das auch der neue Test messen soll. Die Ergebnisse beider Tests werden korreliert.

Der andere Test stellt das „Kriterium“ dar, an dem der neue Test validiert wird. Anstelle eines anderen Tests bzw. zusätzlich kann irgendein Außenkriterium verwandt werden, das in einem sinnvollen Zusammenhang mit dem Merkmal steht, das der neue Test erfaßt. – Man beschreibt dieses Verfahren auch mit dem Begriff der *Übereinstimmungsvalidität* (Parallelen-Validität).

Bei der Konstruktion eines Intelligenztestes wird z. B. überprüft, wie gut die Meßergebnisse dieses Testes mit denen anderer Intelligenztests übereinstimmen. Es wird ferner untersucht, zu welchem Grad die Ergebnisse des Intelligenztestes mit den Leistungen in Schule und Leben übereinstimmen. Hierbei wird angenommen, daß intelligentes Verhalten in einem direkten Verhältnis zum Schul- und Lebenserfolg steht.

Test und Kriterium müssen auf das Gleiche abstellen. Das Kriterium selbst muß valide sein. Beide Forderungen lassen sich in der Regel nur bedingt erfüllen. Aus diesem Grunde sind die bei der Übereinstimmungsvalidität ermittelten Korrelationskoeffizienten meist nicht sehr hoch.

Im Rahmen des Validitätsnachweises für den Grundintelligenztest CFT3 sind Ergebnisse, die Probanden in diesem Test erzielten, mit Ergebnissen korreliert worden, welche die gleichen Probanden in anderen Intelligenztests erzielten. So beträgt z. B. die Korrelation des CFT3 mit dem Intelligenz-Struktur-Test (IST) von Amthauer $r = 0,68$. Die Korrelation mit dem Prüfsystem für Schul- und Bildungsberatung (PSB) von Horn liegt mit $r = 0,66$ bei einem ähnlichen Wert¹⁶.

Wenn Test und Kriterium exakt das Gleiche messen, ist die Korrelation zwischen beiden hoch. Intelligenz-Struktur-Test und Prüfsystem für Schul- und Bildungsberatung prüfen beide, wie der Grundintelligenztest CFT3, die Intelligenz. Der Grundintelligenztest will jedoch nur die allgemeine Grundintelligenz erfassen, während die anderen beiden Verfahren

auf die strukturelle Intelligenz abstellen (Wortflüssigkeit, Technische Begabung, Denkfähigkeit u. a.). Dabei können in die strukturellen Intelligenztests vorbildungsgebundene Einflüsse eingehen. Aus dem Grundintelligenztest ist versucht worden, „intelligenzfremde“ Einflußfaktoren herauszuhalten. Dieser Test erfaßt Intelligenz auf eine andere Weise, und der angezielte Intelligenzbereich ist als sog. Grundintelligenz auch theoretisch anders gefaßt als bei den anderen Verfahren. Die Korrelation kann demzufolge auch nicht ausgesprochen hoch sein.

Psychologische Meßverfahren dienen häufig zur Vorhersage künftigen Verhaltens. Es muß etwas vorhergesagt werden, was zeitlich sehr viel später anfällt als die Testwerte, die man heute erhält. Ein Test hat *Vorhersage-Validität* (prognostische oder prädiktive Validität), wenn seine Vorhersagen in einem hohen Ausmaß mit dem tatsächlich später Eingetretenem übereinstimmen. – Es soll z. B. auch aufgrund eines Intelligenztestes über die Zuweisung von Schülern am Ende der Orientierungsstufe zu den verschiedenen Schulformen entschieden werden. Der einzusetzende Intelligenztest muß eine prognostische Validität haben. Es muß ein Zusammenhang zwischen der Leistung in diesem Test und dem späteren Schulerfolg bestehen.

Beim Grundintelligenztest CFT3 beträgt die Korrelation bei der prognostischen Validität zwischen Testleistung und Schulerfolg (gemessen am Gesamtnotendurchschnitt) im Durchschnitt $r = 0,35$. Die Korrelationskoeffizienten variieren für die einzelnen Unterrichtsfächer und Schulformen. Der höchste Koeffizient liegt bei $r = 0,58$ (Korrelation zwischen Mathematiknote und Leistung im Grundintelligenztest bei Realschulklassen)¹⁷. Wie Sie sehen, ist der Zusammenhang zwischen Intelligenztestleistung und Schulerfolg nicht sehr hoch. Zwei Begründungen lassen sich anführen. (1) Weder die Schulnoten noch der Grundintelligenztest sind valide, wobei nicht gefragt werden soll, welches von beiden weniger valide ist. (2) Die Schulleistung wird nicht allein durch die Intelligenzleistung beeinflusst.

8.4 Konstruktvalidität

Die Frage nach der Konstruktvalidität eines Testes bezieht sich darauf, inwieweit ein Test das abstrakte psychologische Gedankengerüst wieder spiegelt, das seiner Konstruktion zugrundeliegt. Psychologische Konstrukte sind ein Gefüge von angenommenen Variablen, die heuristisch gefunden oder aus bereits bestehenden psychologischen Theorien abgeleitet sind. Angst, technisches Verständnis, Intelligenz, Leistungsmotivation oder kritisches Denken sind z. B. psychologische Konstrukte. Ein Test,

der Angst mißt, ist nur dann „konstrukt valide“, wenn nachgewiesen werden kann, daß die vom Test erfaßten Merkmale in genügender Übereinstimmung mit dem theoretischen Konstrukt Angst stehen.

J. C. Stanley und K. D. Hopkins führen die Entwicklung der Intelligenzmessung als klassisches Beispiel für eine Konstruktvalidierung an¹⁸. Am Anfang versuchte man, die Intelligenz über Reaktionszeitmessungen, Hörgedächtnisprüfungen und andere psychomotorische oder psychophysikalische Messungen zu bestimmen. Dies korrelierte jedoch nicht mit dem, was man als so verstandenes, intelligentes Verhalten beobachten konnte, wie es sich z. B. in den Schulnoten ausdrücken soll. Daraufhin wurden kognitive Testaufgaben konstruiert wie z. B. Zahlenfolgen, aus denen jenes Glied herauszusuchen ist, das nicht in die Gesetzmäßigkeit der Folge hineinpaßt. Die über kognitive Aufgaben gewonnenen Meßergebnisse korrelierten mit anderen „Intelligenzvariablen“ in etwa so, wie man es von einer Intelligenzmessung erwartete.

Nach einem langen Forschungsprozeß erbrachten die Intelligenzteste Meßwerte, die sie sich stark den logischen und theoretischen Erwartungen nähern, u. a.: Die Meßergebnisse korrelieren mit dem Alter der Probanden bis zur Reife und bleiben dann verhältnismäßig konstant, die Intelligenzmeßwerte stehen in einer positiven Beziehung zur Schulleistung, Probanden mit Abnormalitäten (wie z. B. Down's Syndrom¹⁹) erreichen äußerst geringe Leistungen oder die Korrelation bei eineiigen Zwillingen ist hoch. – Die hier genannten Beispiele verdeutlichen den nie endenden Vorgang einer Konstruktvalidierung. Viele Kriterien werden gebraucht, um herauszufinden, was der Test mißt und was er nicht mißt. Die Konstruktvalidierung ist eine Frage des logischen wie auch des empirischen Vorgehens.

8.5 Zusammenfassung

Die Frage nach der Validität eines Testes zielt darauf ab, ob ein Test als Meßinstrument auch tatsächlich das mißt, was er zu messen vorgibt. Drei Klassen von Validität lassen sich unterscheiden. (1) Die *Inhaltsvalidität* ist der relevante Typ für Schulleistungstests; über eine logische Analyse gilt es zu klären, ob die Testaufgaben das Gelehrte widerspiegeln. (2) Die *kriterienbezogene Validität* ist eine Frage des empirischen Vorgehens. Über einen Korrelationskoeffizienten wird der Zusammenhang der Testwerte mit einem plausiblen Kriterium bestimmt. Es wird unterschieden zwischen der Übereinstimmungs- und der Vorhersagevalidität. (3) Die *Konstruktvalidität* bezieht sich darauf, inwieweit der Test das abstrakte psychologische Gedankengerüst widerspiegelt, das seiner Konstruktion zu grundeliegt.

9. Reliabilität

9.1 Der Begriff der Reliabilität (Genauigkeit, Zuverlässigkeit)

Ein Test als Meßinstrument ist reliabel²⁰, wenn er das, was er mißt, genau mißt. Bei der Reliabilität geht es um die Genauigkeit, die Zuverlässigkeit, die Treffsicherheit eines Meßinstrumentes, unabhängig davon, ob der Test das mißt, was er messen soll. Schießt ein Gewehr „Punkt“, d. h. sind alle Einschläge auf einer Zielscheibe dicht um irgend einen Punkt zentriert, ist das Gewehr reliabel. (Liegen diese Einschläge auch in der Mitte der Zielscheibe auf der 10, ist das Gewehr valide.)

Die folgenden Beispiele führen hin auf die Beziehung zwischen Reliabilität und Validität.

Ein Mann fragt nach dem Alter einer ihm noch nicht sehr bekannten Frau. Er erhält „28 Jahre“ zur Auskunft. Nach Wochen fragt er noch einmal und erhält wieder „28 Jahre“ zur Antwort. Nach einem Jahr kann er anhand ihrer Geburtsurkunde feststellen, daß sie tatsächlich 36 Jahre alt ist, zum Zeitpunkt seiner Fragen war sie also 35 Jahre alt. Seine Befragung als Messung war reliabel. Er erhielt zweimal die gleiche Auskunft, d. h. zweimal den gleichen Meßwert. Die Messung war aber nicht valide, da das angegebene Alter nicht mit dem tatsächlichen übereinstimmt.

Nehmen Sie an, der Mann fragt eine zweite Frau nach ihrem Alter. Auch sie gibt wiederholt an, daß sie 28 Jahre alt sei. Ihr angegebenes Alter stimmt auch mit ihrem tatsächlichen überein, wie der Mann nach einer Zeit anhand eines persönlichen Dokuments feststellen kann. In diesem Fall ist die wiederholte Befragung als Messung reliabel und valide. Die Auskunft, d. h. die Meßgröße, war immer die gleiche und stimmte mit dem tatsächlichen überein.

Der Mann möge nun eine dritte Frau nach ihrem Alter fragen. Sie gibt beim ersten Mal 26 Jahre, beim zweiten Mal 28 Jahre an. Die Befragung als Messung ist unreliabel. Bei der Meßwiederholung änderte sich der Meßwert. Damit ist die Messung auch gleichzeitig unvalide.

Sind die Meßwerte genau, besteht die Möglichkeit, daß sie auch gültig sind: Gibt eine Frau zweimal das gleiche Alter an, dann ist es möglich, daß sie das wahre Alter genannt hat. Sind die Meßwerte ungenau, können sie auch zu keiner gültigen Aussage führen. Das wahre Alter der Frau kann 26, 28 oder auch ein anderes Alter sein.

Reliabilität und Validität

Wie oben gezeigt wurde, ist die Reliabilität eine notwendige, aber nicht eine hinreichende Bedingung für die Validität eines Testes²¹. Mit anderen

Worten, wenn die Reliabilität erfüllt ist, muß damit nicht notwendigerweise die Validität gegeben sein.

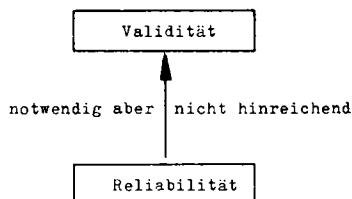


Abb. 22:
Beziehung zwischen Reliabilität und Validität

9.2 Der Standardmeßfehler

Der aktuelle und der wahre Rohwert

Für die folgenden Überlegungen sind die Begriffe „wahrer“ und „aktueller“ Rohwert grundlegend. An einem Gedankenexperiment lassen sich diese Begriffe verdeutlichen. Wenn Sie ein Fünf-D-Markstück an die Decke eines Zimmers werfen und es auf den Boden fallen lassen, wird entweder das Wappen oder die Zahl des Fünf-Markstückes oben liegen. Bei Wappen wird der Rohwert 1 und bei Zahl der Rohwert 0 vergeben. Würden Sie nun hundertmal die Münze in gleicher Weise werfen, erhalten Sie eine Versuchsreihe. Liegt 55mal das Wappen der Münze oben, beträgt die erreichte Rohwertzahl 55. Liegt bei einer zweiten Versuchsreihe von 100 Würfeln 45mal das Wappen der Münze oben, beträgt die Rohwertzahl 45. Die Rohwertzahlen 45 und 55 sind die jeweils bei einer Meßreihe erreichten Rohwerte, die sog. „aktuellen“ Rohwerte.

Die „wahre“ Leistung im Münzenwurf wird sich ermitteln lassen, wenn die Versuchsreihen oft wiederholt werden und die Rohwertzahlen gemittelt werden. Würden Sie hundertmal Reihen mit je 100 Würfeln durchführen, so würden Sie feststellen, daß die erreichten „aktuellen“ Rohwertzahlen zwischen 35 und 65 liegen und daß die Rohwertzahl 50 am häufigsten vorkommt. Fünfzig wird auch der Mittelwert der erreichten Rohwertzahlen sein²².

Wenn Sie bei der ersten Versuchsreihe eine Rohwertzahl von 55 erreichten, so war das „Glück“. Sie erreichten fünf Rohwerte mehr als nach dem Zufall im Mittel zu erwarten gewesen wäre. Bei einer Rohwertzahl von 45 hätten Sie „Pech“ gehabt. Sie lagen 5 Werte unter dem zu erwartenden Wert von 50. Pech und Glück werden sich bei einer sehr oft wiederholten Messung ausgleichen, so daß sich die „wahre“ Leistung bei 50, dem Mittelwert aller Rohwertzahlen, einstellt.

Definition 19:

Der „aktuelle“ Rohwert ist der bei einer Messung ermittelte Rohwert. Der „wahre“ Rohwert ist der Wert, auf den sich die aktuellen Rohwerte bei wiederholter Messung einpegeln: Mittelwert der aktuellen Rohwerte.

Zur Ableitung des Meßfehlerbegriffes werden im folgenden die beiden Begriffe aktueller und wahrer Rohwert auf einen Test übertragen.

Der Meßfehler

Es soll der Wortschatz getestet werden, den Schüler eines Englischkurses am Schuljahresende besitzen. Dazu wird aus einem englisch-deutschen Wörterbuch eine Zufallsstichprobe von 100 Wörtern gezogen. Diese Wörter werden von den Schülern ins Deutsche übersetzt. Es ist nun von Interesse, wieviel Prozent der 100 Wörter ein Schüler übersetzen kann. Das eigentliche Interesse muß aber darauf gerichtet sein, festzustellen, wie groß der Prozentsatz der bekannten Wörter in der Population der Wörter ist. Wird für einen zweiten Test eine neue Zufallsstichprobe von 100 Wörtern gezogen, erreichen vermutlich wenige Schüler (wenn überhaupt welche) die gleiche Leistung, die sie auch in der ersten Testform erzielten.

Einige Schüler haben mit der ersten Testfassung Glück. Sie finden mehr Wörter vor, die sie kennen. Andere haben Glück mit der zweiten Testfassung. Sie haben in der ersten Fassung weniger bekannte Wörter angetroffen. Von Interesse ist nun, wie groß der Prozentsatz der bekannten Wörter eines Schülers im Schnitt bei möglichst vielen Testfassungen ist (theoretisch bei unendlich vielen). Dieser Durchschnittswert wäre die „wahre“ Leistung im Wortschatztest. Sie ist frei von dem Glück und dem Pech, das von der Wörterwahl einer Testform abhängt. Die wahre Leistung ist ferner frei von anderen Zufallsgrößen, wie z.B. dem Glück im Raten oder der momentanen psychischen Konstellation.

Nehmen Sie nun an, der Schüler Günther S. hat sich der Mühe unterzogen, immer wieder neue 100 Wortstichproben aus dem englisch-deutschen Wörterbuch zu übersetzen. In Testfassung A erreichte er eine Leistung von 72 richtig übersetzten Wörtern. In Fassung B 66, in C 71, in D 74, in E 69 usw. Seine erreichten „aktuellen“ Rohwerte mögen normalverteilt sein und sich um einen Mittelwert von 70 zentrieren (Abbildung 23).

Die wahre Leistung von Günther S. liegt bei 70, dem Mittelwert der aktuellen Rohwerte. Alle Abweichungen von dem Rohwert 70 sind nun die *Fehler der Messung*, verursacht durch Zufallseinflüsse. In der Fassung A (72 Rohwerte) beträgt der Fehler der Messung 2 Rohwerte. Manchmal

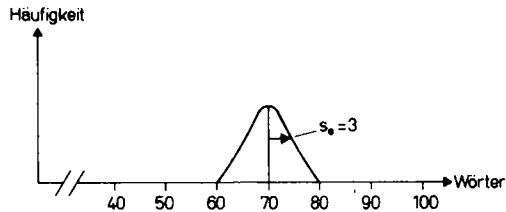


Abb. 23:

Die Verteilung der Übersetzungsleistung vom Englischen ins Deutsche bei einem Schüler für wiederholt neue 100 Wortstichproben aus dem Wörterbuch.

war der Meßfehler groß, in der Regel war er aber klein (vgl. Abbildung 23). Je kleiner der Meßfehler, um so größer ist die Reliabilität, d. h. die Genauigkeit der Messung.

Jeder erreichte, „aktuelle“ Rohwert besteht aus zwei Komponenten, dem „wahren“ Rohwert und dem durch Zufall bedingten Fehlerwert der Messung, dem Meßfehler. Bei der Testfassung A setzt sich der aktuelle Rohwert 72 aus dem wahren Rohwert 70 plus dem Meßfehler von 2 Rohwerten zusammen. Bei der Fassung B ergibt sich die aktuelle Leistung von 66 aus der wahren Leistung 70 minus der Fehlerleistung 4. Der Meßfehler ist hier negativ (-4).

Es ist:

$$\text{Aktueller Rohwert} = \text{Wahrer Rohwert} + \text{Meßfehler}$$

Der Meßfehler läßt sich aus der Differenz zwischen aktuellem und wahrem Rohwert definieren.

Definition 20:

Der *Meßfehler* für einen Probanden bei einmaliger Testdurchführung besteht aus der Differenz zwischen aktuellem und wahrem Rohwert.

Meßfehler = Aktueller Rohwert minus Wahrer Rohwert.

Der Standardmeßfehler

Für die Verteilung der aktuellen Rohwerte beim Schüler Günther S. (Abbildung 23) läßt sich die Standardabweichung berechnen. Diese Standardabweichung wird als der Standardmeßfehler (s_e) bezeichnet. Der Index e steht für das englische Wort „error“. Der Standardmeßfehler ist ein Maß für die Diskrepanz zwischen aktuellen und wahren Rohwerten und damit ein wichtiges Kennmaß für Tests. In Abbildung 23 ist $s_e = 3$, d. h. für den Englisch-Wortschatztest beträgt der Standardmeßfehler 3 Einheiten (hier wird ein Schüler für alle Schüler gesetzt).

Konnte z.B. ein Schüler 56 Wörter richtig übersetzen, so wird mit 68%iger Wahrscheinlichkeit seine wahre Leistung im Bereich von 56 ± 3 Wörter liegen ($x \pm 1 \cdot s_e$).

Auf die Bedeutung des Standardmeßfehlers wird später noch einmal eingegangen. Vorerst ist abzuleiten, wie der Standardmeßfehler praktikabel gewonnen werden kann. So wie oben vorgegangen wurde, ist ein Schüler für alle Schüler gesetzt worden. Dieser eine Schüler darf bei wiederholter Messung unter exakt gleichen Testbedingungen keinen Lernfortschritt durchmachen. Außerdem bedarf es immer wieder neuer Parallelformen eines Testes. Zweifellos ist es somit wenig praktikabel – wenn überhaupt möglich – einen Schüler endlos „durchzutesten“, um damit eine Verteilung seiner aktuellen Rohwerte zu erhalten, deren Standardabweichung dann gleich dem Standardmeßfehler des Testes ist. Es muß eine Ableitung für den Standardmeßfehler gefunden werden, die von der Standardabweichung der aktuellen Rohwerte aller Personen bei einmaliger Testdurchführung ausgeht.

Bei einer Testdurchführung setzen sich die aktuellen Rohwerte jeder getesteten Person aus zwei Komponenten zusammen:

Aktueller Rohwert = Wahrer Rohwert + Meßfehler.

Das heißt, sind mehrere Personen getestet und wird die Verteilungskurve der aktuellen Rohwerte (x) gezeichnet, setzt sich jeder Punkt dieser Kurve aus zwei Komponenten zusammen: dem wahren Rohwert und dem Fehlerrohwert (Meßfehler). Eine Verteilungskurve wird charakterisiert durch die Standardabweichung s_x oder durch das Quadrat der Standardabweichung s_x^2 , die Varianz. Wenn sich die aktuellen Rohwerte aus wahren Rohwerten und Fehlerrohwerten zusammensetzen, so wird sich auch die aktuelle Varianz aus wahrer Varianz (s_w^2) und Meßfehlervarianz (s_e^2) zusammensetzen:

$$s_x^2 = s_w^2 + s_e^2 \quad (1)$$

Die Genauigkeit, die Reliabilität eines Testes wird um so größer sein, je stärker die aktuelle Varianz (s_x^2) bei einer Testdurchführung mit der wahren Varianz (s_w^2) übereinstimmt.

Über einen Quotienten aus diesen beiden Größen läßt sich die Reliabilität (r) als ein Koeffizient angeben:

$$r = \frac{s_w^2}{s_x^2} \quad (2)$$

Der Koeffizient r in Gleichung (2) wird immer Werte kleiner als 1 annehmen. Wird Gleichung (1) in Gleichung (2) eingesetzt, zeigt sich nämlich, daß mit

$$r = \frac{s_w^2}{s_w^2 + s_e^2} \quad (3)$$

der Nenner dieses Ausdrucks immer größer ist als der Zähler; es sei denn $s_e^2 = 0$. Ist die Fehlervarianz Null, nimmt der Reliabilitätskoeffizient den Wert 1 an. In diesem Falle stimmt die aktuell erhaltene Varianz mit der wahren Varianz überein. Dies dürfte aber wohl nie ganz der Fall sein; eine Fehlervarianz bleibt immer und der Reliabilitätskoeffizient r ist kleiner als 1.

Aus Gleichung (1) und (2) läßt sich nun durch algebraische Umformungen der Standardmeßfehler s_e ermitteln:

Aus (1) folgt für s_w^2 :

$$s_w^2 = s_x^2 - s_e^2 \quad (4)$$

Gleichung (4) wird in Gleichung (2) eingesetzt:

$$r = \frac{s_x^2 - s_e^2}{s_x^2} \quad (5)$$

$$r = \frac{s_x^2}{s_x^2} - \frac{s_e^2}{s_x^2}$$

$$r = 1 - \frac{s_e^2}{s_x^2}$$

Der Reliabilitätskoeffizient – und damit die Reliabilität – ist Null, wenn die Fehlervarianz (s_e^2) genauso groß ist wie die aktuelle Rohwertvarianz (s_x^2). In diesem Falle ist mit $s_e^2 = s_x^2$:

$$r = 1 - \frac{s_x^2}{s_x^2} = 1 - 1 = 0.$$

Sind die Standardabweichung s_x der aktuellen Rohwertverteilung und der Standardmeßfehler s_e gleich groß, bedeutet dies, daß die aktuellen Rohwertunterschiede von Person zu Person alle zu Lasten der Meßfehler gehen.

Ist die Reliabilität perfekt, gehen alle aktuellen Rohwertunterschiede zu Lasten der Unterschiede in den wahren Rohwerten. Der Standardmeßfehler ist in diesem Falle Null und der Reliabilitätskoeffizient ist 1.

$$(r = 1 - \frac{0}{s_x^2} = 1 - 0 = 1.)$$

Die Auflösung von Gleichung (5) nach s_e ergibt nun:

$$\begin{array}{ll}
 r = \frac{s_x^2 - s_e^2}{s_x^2} & \cdot s_x^2 \\
 r \cdot s_x^2 = s_x^2 - s_e^2 & s_x^2 \text{ unter Vorzeichenwechsel auf} \\
 & \text{die linke Seite bringen} \\
 -s_x^2 + r \cdot s_x^2 = -s_e^2 & \cdot (-1) \\
 s_x^2 - r \cdot s_x^2 = s_e^2 & s_x^2 \text{ ausklammern} \\
 s_x^2 (1 - r) = s_e^2 & \text{"Wurzel ziehen" auf beiden Seiten} \\
 \sqrt{s_x^2 (1 - r)} = \sqrt{s_e^2} & \\
 s_x \sqrt{(1 - r)} = s_e & \text{Seiten vertauschen} \\
 \boxed{s_e = s_x \sqrt{(1 - r)}} &
 \end{array}$$

Der Standardmeßfehler s_e eines Testes ist eine Funktion der Standardabweichung der aktuellen Rohwerte in dem Test und seinem Reliabilitätskoeffizienten.

Definition 21:

Der Standardmeßfehler (s_e) eines Testes ist eine Funktion der Standardabweichung der aktuellen Rohwerte (s_x) und des Reliabilitätskoeffizienten (r). Der Standardmeßfehler bestimmt sich über:

$$s_e = s_x \sqrt{(1 - r)} .$$

9.3 Die Bestimmung des Reliabilitätskoeffizienten

Der Standardmeßfehler, wie er oben abgeleitet wurde, enthält nicht mehr den unbestimmbaren Ausdruck der wahren Varianz (s_w^2). Dafür tritt ein neuer Ausdruck auf: der Reliabilitätskoeffizient r . Im folgenden wird aufgezeigt, welche empirischen Wege es gibt, diesen Reliabilitätskoeffizienten zu bestimmen. Dazu folgende Vorüberlegungen:

Unzuverlässige Meßergebnisse ergeben sich bei Tests aufgrund a) der mangelnden Präzision des Meßinstrumentes und b) aufgrund dessen, daß der Meßgegenstand sich rasch verändern kann.

Zu a)

Wenn z.B. die Intelligenz gemessen werden soll, gibt es theoretisch, wie man sagt, ein „Universum“ von Testaufgaben, die eine Aussage über die Intelligenz machen können. Jeder Intelligenztest kann aber nur eine Auswahl von Testaufgaben vornehmen. Für den einen Probanden kann die Auswahl glücklich sein, für den anderen weniger glücklich. Werden zwei gleich große Auswahlen für einen Intelligenztest getroffen (Form A und Form B) müßte ein Proband in beiden Formen die gleiche Leistung erzielen, wollte Form A oder Form B des Intelligenztestes exakt präzise messen. Dies wird aber nie ganz der Fall sein. Bei jeder Auswahl von Testaufgaben werden immer auch solche Aufgaben vorkommen, die der Proband – aus welchen Gründen auch immer – leichter oder schwerer bearbeiten kann als Aufgaben bei einer anderen Auswahl.

In Abbildung 24 soll der große Kreis, der die kleinen Kreise umschließt, das Universum der Testaufgaben darstellen, die für ein psychologisches Merkmal, wie z.B. die Intelligenz, relevant sind. In die Testform A gelangen nur die mit A, in die Testform B nur die mit B gekennzeichneten Testaufgaben. – Die Auswahl der Aufgaben beeinflusst die Präzision, d. h. die Reliabilität eines Testes. Diese Fehlerquelle eines Testes ließe sich nur ausschalten, wenn alle Aufgaben des Aufgabenuniversums in einem Test aufgenommen werden könnten.

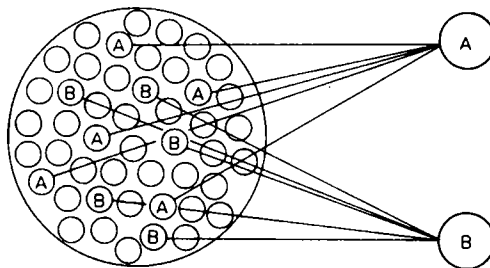


Abb. 24:

Die Auswahl der Aufgaben für Testformen aus dem Universum der Testaufgaben.

Zu b):

Ungenauere Meßergebnisse lassen sich neben der mangelnden Präzision des Meßinstrumentes auch auf den Meßgegenstand zurückführen. Dies trifft zu, wenn der Gegenstand, der gemessen werden soll, sich rasch ver-

ändert. Solche instabilen Persönlichkeitsmerkmale können z. B. sein: die Stimmung, der Neurotizismus, die Angst. Wenn die Stabilität des zu messenden Merkmals gering ist, wird das einzelne Meßergebnis unzuverlässig. Die punktuell vorgenommene Messung gibt nur eine recht ungenaue Abschätzung der Ausprägung des zu messenden Merkmals wieder.

Hinzu kommt die Bedingungsinkonstanz der Testsituation. Eine Person mag heute mit einer ganz bestimmten psychischen Konstellation (z. B. körperliches Wohlbefinden) in eine Testsituation hineingehen. Morgen, wenn der Test wiederholt wird, kann sich die Konstellation für diese Person schon wieder geändert haben. Die Messung wird demzufolge unrelia-
bel.

Die Reliabilität eines Testes wird durch verschiedene Fehlergrößen bestimmt. Die operationalen Zugänge zur Bestimmung der Reliabilität, wie sie im folgenden aufgezeigt werden, können immer nur bestimmte Fehlergrößen in Rechnung stellen. So gesehen existiert nicht eine Reliabilität eines Testes, sondern es existieren immer nur verschiedene methodische Zugänge zur Reliabilität. Diese Zugänge können

- a) bei einmaliger Testgabe (eine Messung) oder
- b) bei zweimaliger Testgabe (zwei Messungen) gefunden werden.

Zu a)

9.31 Testhalbierungs-Reliabilität (*Split-half-Reliabilität*)

Der gesamte Test wird bei mehreren Personen nur einmal durchgeführt. Danach wird er in zwei Hälften aufgeteilt. Jede Hälfte wird als Paralleltest zur anderen Hälfte aufgefaßt. Die Halbierung geschieht meist so, daß wechselweise eine Testaufgabe der einen Hälfte zugeordnet wird und die nächste der anderen, z. B. die ungeradzahligen Testaufgaben 1, 3, 5, 7... in die Testhälfte 1 und die geradzahligen Aufgaben 2, 4, 6, 8... in die Hälfte 2. Beide Testhälften werden korreliert. Dies ergibt den Korrelationskoeffizienten r_{12} . Über die Spearman-Brown Formel läßt sich, ausgehend von dem Korrelationskoeffizienten der beiden Testhälften, der Reliabilitätskoeffizient (gleich die Reliabilität) für den gesamten Test aus Testhälfte 1 und 2 schätzen.

Definition 22:

*Testhalbierungs-Reliabilität: Spearman-Brown Formel*²³:

Ausgehend von dem Korrelationskoeffizienten der beiden Testhälften r_{12} wird die Reliabilität des gesamten Testes geschätzt über:

$$r_{tt} = \frac{2 \cdot r_{12}}{1 + r_{12}}$$

mit r_{tt} Reliabilitätskoeffizient,
 r_{12} Korrelationskoeffizient für die Korrelation
 der Testhälfte 1 mit der Testhälfte 2.

Für den Reliabilitätskoeffizienten r wird im Index – aus dem Englischen übernommen – tt (test, test) geschrieben. Dies soll andeuten, daß zwei Tests miteinander korreliert werden. Bei der Testhalbierungs-Reliabilität sind es die zwei Testhälften 1 und 2, die miteinander korreliert werden. (– Ebenso wird beim Validitätskoeffizienten r_{tc} geschrieben. Hier wird (englisch) test mit (englisch) criterion korreliert.)

Der Spearman-Brown Formel liegt die Annahme zugrunde, daß die Standardabweichungen der zwei Testhälften 1 und 2 gleich groß sind ($s_1 = s_2$).

Einfacher zur Bestimmung der Testhalbierungsreliabilität ist die Formel von Flanagan. Der Korrelationskoeffizient für die beiden Testhälften braucht nicht mehr berechnet zu werden. Auch wird nicht verlangt, daß die Standardabweichung beider Testhälften gleich groß ist.

Definition 23: -

*Testhalbierungs-Reliabilität: Flanagan-Formel*²⁴:

Die Reliabilität wird ausgehend von den Standardabweichungen der Testhälften 1 und 2 und der Standardabweichung des gesamten Testes berechnet über:

$$r_{tt} = 2 \left(1 - \frac{s_1^2 + s_2^2}{s_x^2} \right)$$

mit r_{tt} Reliabilitätskoeffizient,
 s_1, s_2 Standardabweichung der Rohwerte
 in Testhälfte 1 und in Testhälfte 2
 s_x Standardabweichung der Rohwerte
 des gesamten Testes.

Für den Grundintelligenztest CFT 3 Skala 3 werden bei der Testhalbierungsreliabilität für drei verschiedene Stichproben von Probanden Reliabilitätskoeffizienten r_{tt} von 0,82, 0,91 und 0,95 angegeben²⁵. Bei der Drahtbiegeprobe (D-B-P) wurden Testhalbierungskoeffizienten von $r_{tt} = 0,91$ und $r_{tt} = 0,88$ erzielt²⁶.

Die Testhalbierungsreliabilität wird in der Regel bei standardisierten psychologischen Meßverfahren sehr hoch sein. Bei einem Intelligenztest sollte ein Wert um 0,90 und besser verlangt werden.

9.32 Konsistenzanalyse

Die Konsistenzanalyse ist als eine Verallgemeinerung der Testhalbierungsmethode anzusehen. Auch sie setzt nur eine einmalige Testdarbietung voraus. Ihr liegt der Gedanke zugrunde, daß ein Test nicht nur in zwei vergleichbare Hälften sondern in gleich viele Teile aufgespalten werden kann, wie er Aufgaben enthält. Ebenso wie die Konsistenz zweier Testhälften geprüft wird, kann auch die Konsistenz der einzelnen Aufgaben geprüft werden: 1) Jede Aufgabe wird mit jeder anderen korreliert (interkorreliert), 2) der mittlere Korrelationskoeffizient dieser Koeffizienten wird als der Reliabilitätskoeffizient für die typische Aufgabe des gesamten Tests angesehen, 3) mit einer abgewandelten Spearman-Brown Formel läßt sich der Reliabilitätskoeffizient für den gesamten Test schätzen.

Wollte man nach dem oben genannten Verfahren die Reliabilität berechnen, müßten bereits bei 10 Aufgaben in einem Test eine Unmenge

von Interkorrelationen berechnet werden ($\frac{k(k-1)}{2} = \frac{10 \cdot 9}{2} = 45$ In-

terkorrelationen). Es wurden daher Formeln entwickelt, mit der die Konsistenz eines Testes auf der Grundlage des o.g. Verfahrens geschätzt werden kann. Bekannt sind hier die Kuder-Richardson Formeln 20 und 21. Kuder und Richardson numerierten die von ihnen vorgeschlagenen Formeln durch. Die Formeln 20 und 21 wurden später vielfach verwendet und gingen so in die Literatur ein.

Definition 24:

Konsistenzanalyse: Kuder-Richardson Formel 20²⁷:

Über die Schwierigkeit der Testaufgaben, die Standardabweichung des gesamten Testes und die Zahl der Testaufgaben, berechnet sich der Reliabilitätskoeffizient über:

$$r_{tt} = \frac{k}{(k-1)} \cdot \left[1 - \frac{\sum p \cdot q}{s_x^2} \right]$$

mit

- r_{tt} Reliabilitätskoeffizient,
- k Zahl der Testaufgaben,
- p Schwierigkeit der einzelnen Testaufgabe:
Anteil der Probanden, die die jeweilige Aufgabe richtig lösten,
- q Anteil der Probanden, die die jeweilige Aufgabe nicht richtig lösten ($q = 1 - p$),

s_x Standardabweichung der Rohwerte im gesamten Test,
 $\Sigma p \cdot q$ Summe der Rohwerte $p \cdot q$ über alle Aufgaben.

Bisher noch nicht vorgestellt ist die Größe p , die Schwierigkeit einer Testaufgabe, und damit auch $q = 1 - p$. p als Schwierigkeitsindex gibt den Anteil der Probanden an, welche die jeweilige Aufgabe richtig lösten. Haben in einem Test die Aufgabe a 15 von 30 Schülern richtig gelöst, dann beträgt die Schwierigkeit p dieser Aufgabe $p = \frac{15}{30} = 0,5$. Entsprechend ist $q = 1 - p = 1 - 0,5 = 0,5$. Lösen nur 8 von 30 die Aufgabe b, dann ist diese Aufgabe schwer: $p = \frac{8}{30} = 0,27$, $q = 1 - 0,27 = 0,73$. Konnten 28 von 30 die Aufgabe c richtig beantworten, ist diese Aufgabe leicht: $p = \frac{28}{30} = 0,93$, $q = 0,07$. Je größer der Schwierigkeitsindex ist, um so leichter ist die Aufgabe. Je kleiner der Index ist, desto schwerer ist die Aufgabe.

Zur Berechnung der Kuder-Richardson Formel 20 muß nun zu jeder Testaufgabe der Schwierigkeitsindex p sowie $q = 1 - p$ ermittelt werden. Für jede Aufgabe sind beide Größen zu multiplizieren und diese Produkte anschließend aufzusummieren, um den Ausdruck $\Sigma p \cdot q$ zu erhalten. Dies ist ein wenig mühselig. Eine Abschätzung des Ausdrucks $\Sigma p \cdot q$ über den arithmetischen Mittelwert der Rohwerte des gesamten Testes $\bar{x}$ führte zur Kuder-Richardson Formel 21.

Definition 25:

*Konsistenzanalyse: Kuder-Richardson-Formel 21*²⁸:

Ausgehend von der Zahl der Testaufgaben, der Standardabweichung und dem Mittelwert des gesamten Testes berechnet sich die Reliabilität über:

$$r_{tt} = \frac{k \cdot s_x^2 - \bar{x}(k - \bar{x})}{(k - 1) s_x^2}$$

mit

r_{tt} Reliabilitätskoeffizient,
 k Anzahl der Testaufgaben,
 s_x Standardabweichung der Rohwerte des gesamten Testes,
 $\bar{x}$ Arithmetischer Mittelwert der Rohwerte des gesamten Testes.

Die Abschätzung des Ausdruckes $\Sigma p \cdot q$, die von der Kuder-Richardson Formel 20 zur Formel 21 führt, macht für Formel 21 eine Einschränkung notwendig. Die Schwierigkeit p der Testaufgaben muß in etwa gleich sein, wenn nicht die Formel 21 zu einem niedrigeren Ergebnis als Formel 20 führen soll. Variiert der Schwierigkeitsindex der meisten Aufgaben zwischen $p = 0,30$ und $p = 0,70$, ist der Fehler gegenüber der Formel 21 jedoch noch geringfügig²⁹. Es dürfen somit nur wenige Testaufgaben extrem leicht oder extrem schwer sein.

Die Kuder-Richardson Formel 21 ist eine Faustformel, die eine grobe Konsistenzschätzung eines Testes bereits dann abgibt, wenn die Elementarstatistik-Werte s_x und $\bar{x}$ vorliegen. Es wird aber kein Fehler begangen, wenn diese Formel zur Reliabilitätsbestimmung verwendet wird. Eine Aufgabenanalyse deckt auf, ob die Annahme des etwa gleichen Schwierigkeitsgrades der Testaufgaben zu stark verletzt wird, wie es bei einem vom Lehrer selbst konstruierten Test leicht der Fall sein kann. Will man aber schon vor der Aufgabenanalyse eine Abschätzung der Reliabilität vornehmen, führt die Kuder-Richardson Formel 21 nicht zu einer fälschlich als zu hoch ausgewiesenen Reliabilität. Eher das Gegenteil kann der Fall sein. Gegenüber der korrekteren Formel 20 wird mit der Formel 21 die Reliabilität immer niedriger eingeschätzt.

Im Test Vitamine und Mineralstoffe, der eingangs vorgestellt wurde, beträgt die Zahl der Aufgaben $k = 20$, die Standardabweichung $s_x = 3,00$ und der Mittelwert $\bar{x} = 12$. Der Reliabilitätskoeffizient für diesen Test beträgt demnach, geschätzt nach Kuder-Richardson Formel 21:

$$r_{tt} = \frac{20 \cdot (3^2) - 12(20 - 12)}{(20 - 1) \cdot (3^2)}$$

$$= 0,49.$$

Der Reliabilitätskoeffizient ist nicht hoch. Es handelt sich noch um die erste Testfassung. Inwieweit die Annahme der Kuder-Richardson Formel 21 bezüglich des in etwa gleichen Schwierigkeitsgrades verletzt worden ist und so ein ungerechtfertigt niedriger r_{tt} ausgewiesen worden ist, wird die Aufgabenanalyse noch zeigen (siehe unten).

Um übliche Werte bei standardisierten, erwerbbaren Tests anzugeben: für die Drahtbiegeprobe (D-B-P) von G. A. Lienert wurden Koeffizienten von $r = 0,87$ und $r = 0,89$ bei der Konsistenzanalyse ermittelt³⁰.

Zu b)

9.33 *Paralleltest-Reliabilität*

Bei den bisher vorgestellten Verfahren der Reliabilitätsbestimmung (Testhalbierungs-Reliabilität, Konsistenzanalyse) brauchte der Test immer nur einmal dargeboten zu werden. Auf der Testwiederholung beruht das Verfahren der Paralleltest-Reliabilität. Hier werden ein und denselben Probanden zwei einander stark ähnelnde Tests vorgelegt und deren Ergebnisse miteinander korreliert. Der Korrelationskoeffizient ergibt den Reliabilitätskoeffizienten. Je genauer das Ergebnis der ersten Testdurchführung mit dem Ergebnis der zweiten Testdurchführung übereinstimmt, um so größer ist die Reliabilität des Testes.

Parallelformen eines Testes sind schwierig zu entwickeln. Für Leistungstests in der Schule, wie z. B. dem informellen Leistungstest über Vitamine und Mineralstoffe dürfte nur schwerlich ein Paralleltest zu konstruieren sein. Die Testaufgaben in einem Schulleistungstest haben in der Regel einen gewissen Einmaligkeitscharakter. Darüber hinaus können die Lösungskonzepte aus dem ersten Test auf den zweiten übertragen werden, so daß von dem ersten zum zweiten Test ein Lern- oder Übungsfortschritt zu erwarten ist. Die Paralleltest-Reliabilität kommt somit für Schulleistungstests weniger in Frage. Bei diesen Tests wird man sich für die Reliabilitätsbestimmung der Testhalbierungsmethode oder der Konsistenzanalyse bedienen; also Verfahren der einmaligen Testdarbietung.

Die Schwierigkeiten, einen streng vergleichbaren Paralleltest zu entwickeln, lassen erwarten, daß die Korrelation zwischen zwei parallelen Testformen nicht ausgesprochen hoch sein wird. Gute Tests erreichen eine Paralleltest-Reliabilität von $r = 0,85$ und höher. Ein $r = 0,80$ kann bereits als zufriedenstellend angesehen werden.

9.34 *Retest-Reliabilität*

Bei der Retest-Reliabilität wird der gleiche Test denselben Probanden zu zwei verschiedenen Zeitpunkten vorgegeben. Die Ergebnisse der ersten und zweiten Messung werden korreliert. Der Korrelationskoeffizient ergibt den Reliabilitätskoeffizienten. Ein guter Test müßte bei denselben Probanden heute und in vier Wochen zu möglichst gleichen Ergebnissen kommen.

Die Zeitpunkte der ersten und zweiten Messung müssen weit genug auseinanderliegen, um Erinnerungseffekte auszuschließen, die die Reliabilität fälschlich erhöhen. Man spricht in diesem Zusammenhang von einer Scheinreliabilität, die entstehen kann. Die Erinnerungseffekte werden um so größer sein, je kürzer der Abstand zwischen dem Test und seine

Wiederholung ist, und je leichter die Aufgaben behalten werden können (Zahl der Aufgaben, Charakteristika der Aufgaben). Auf der anderen Seite setzt aber das Verfahren der Testwiederholungsreliabilität voraus, daß das zu messende Persönlichkeitsmerkmal stabil bleibt. Der Zeitpunkt der Wiederholungsmessung darf also nicht weit von dem der Erstmessung entfernt sein.

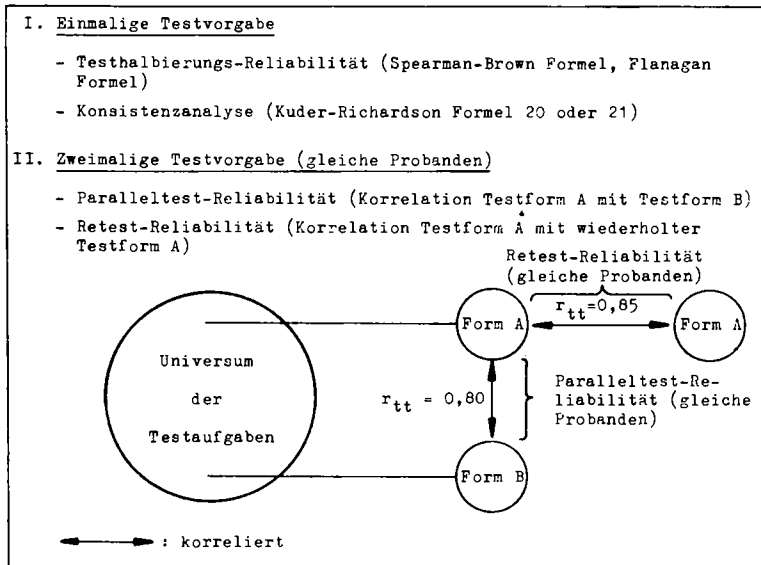
Die Retest-Reliabilität eines Testes wird immer höher sein als seine Paralleltest-Reliabilität. Bei der Wiederholung ein und desselben Tests ist der Testinhalt der gleiche. Es wird nicht eine neue Stichprobe von Testaufgaben aus dem Aufgabenuniversum den Probanden dargeboten, wie dies bei der Reliabilitätsbestimmung über Paralleltests der Fall ist. Während bei der Paralleltest-Reliabilität ein $r_{tt} = 0,80$ und höher verlangt wird, kann bei der Retest-Reliabilität ein $r_{tt} = 0,85$ und höher gefordert werden.

Für den Grundintelligenztest CFT3 Skala 3 liegt die Retest-Reliabilität r_{tt} ohne Zeitverschiebung zwischen Werten von 0,84 bis 0,94. Für ein großes Zeitintervall (mehrere Wochen) schwanken die Werte zwischen $r_{tt} = 0,70$ und $r_{tt} = 0,80$ ³¹. Bei der Drahtbiegeprobe liegen die Retest-Reliabilitätskoeffizienten zwischen $r_{tt} = 0,78$ und $r_{tt} = 0,83$ ³². Hier kann sich ein für die Probanden unterschiedlicher Übungseffekt bemerkbar gemacht haben, der die Wiederholungsreliabilität beeinflusst: Dem einen oder anderen Probanden fällt das wiederholte Biegen des Drahtes leichter. Die Ergebnisse aus der ersten Drahtbiegeprobe korrelieren mit dem Ergebnis aus der zweiten Probe demzufolge niedriger als gefordert.

Die Retest-Reliabilität eignet sich nicht zur Reliabilitätsbestimmung für Schulleistungstests. Die Testpersonen können sich bei einer Wiederholungsmessung mit dem gleichen Test an bestimmte Aufgabenlösungen erinnern, ohne den eigentlichen Lösungsprozeß zu vollziehen.

9.35 Zusammenfassung

Die Verfahren der Retest- und Paralleltestreliabilität erfordern eine zweimalige Testdurchführung. Dies kann in der praktischen Durchführung auf Schwierigkeiten stoßen. Aus diesem Grunde sind Methoden der Reliabilitätsbestimmung, die nur eine einmalige Testgabe erfordern, für von Lehrer selbst konstruierte, informelle Leistungstests eher geeignet – Tabelle 27 gibt einen Überblick über die besprochenen Reliabilitätsarten wieder.



Tab. 27:
Überblick über die Reliabilitätsarten.

9.4 Möglichkeiten der Reliabilitätsverbesserung

Um eine höhere Reliabilität zu erzielen, muß bei der Testrevision auf die Faktoren eingewirkt werden, welche die Reliabilität eines Testes beeinflussen. Drei wesentliche Faktoren, auf die hier eingegangen werden soll, sind

- a) die Testlänge,
- b) die Homogenität der Testaufgaben und
- c) die Trennschärfe bzw. die Schwierigkeit der Aufgaben.

Zu a)

Je länger ein Test ist, desto reliabler d. h. meßgenauer wird er. Erinnern Sie sich an den Englisch-Wortschatztest. Einen Unterschied für die Meßgenauigkeit kann es machen, wenn statt der hier vorgeschlagenen 100 Wörter nur 50 für die Testform gezogen würden. Je weniger Aufgaben aus dem Universum der Aufgaben gezogen werden, desto größer wird der Einfluß des Zufalls. Bei einer geringen Aufgabenzahl mag man gerade auf jene Aufgaben stoßen, die man leichter bzw. schwerer lösen kann, aus welchen Gründen auch immer. Zweifellos aber wird es kaum einen Unterschied machen, wenn Sie für eine Testform 500 statt 600 englische Wörter dem Zufall nach aus dem Wörterbuch herausziehen.

Enthält ein vom Lehrer selbst entwickelter Leistungstest 20 Aufgaben, so ist zu überlegen, ob nicht die Aufgabenzahl um 5 weitere Aufgaben auf 25 erhöht werden kann, um die Reliabilität – sofern nötig – zu verbessern.

Zu b)

Je homogener die Testaufgaben sind, desto reliabler wird der gesamte Test. Ein 50 Aufgaben umfassender Mathematiktest wird reliabler sein als ein 50 Aufgaben umfassender Leistungstest, der den Englisch-, Deutsch-, Physik- und Mathematikunterricht abdeckt. Erfasst ein Test mehr als einen Aspekt, müssen aus mehr als einem Testuniversum Aufgaben gezogen werden. Jeder Aufgabenbereich wird von verhältnismäßig wenigen Aufgaben vertreten; damit wird die Meßgenauigkeit für die einzelnen Aspekte reduziert. – Je homogener die Testaufgaben sind, desto konzentrierter kann die Messung erfolgen.

In naturwissenschaftlichen Lehrfächern wird sich die Inhaltshomogenität der Testaufgaben leichter erstellen lassen als bei den nicht naturwissenschaftlichen Fächern. Ein Mathematiktest über die quadratische Funktion und über die Lösung quadratischer Gleichungen läßt sich vom Inhalt her homogener gestalten als ein Rechtschreibtest.

Zu c)

Je trennschärfer die Testaufgaben in einem Test sind, desto reliabler wird der gesamte Test sein. Voraussetzung ist die mittlere Schwierigkeit der Aufgaben. Eine Aufgabe, die von der Hälfte aller Testbearbeiter richtig gelöst wird, kann optimal zur Trennung zwischen den Leistungsstarken und Leistungsschwachen und damit zur Meßgenauigkeit beitragen. Der Testentwickler muß die nicht trennscharfen Aufgaben aufspüren und sie durch trennscharfe ersetzen. Wie dies geschieht, wird in dem Kapitel über die Aufgabenanalyse noch ausführlich erläutert.

9.5 Der Standardmeßfehler und sein Vertrauensintervall

Sobald die Reliabilität eines Testes bekannt ist, läßt sich sein Standardmeßfehler (s_e) berechnen. Es ist (Def. 21) $s_e = s_x \sqrt{1-r}$, mit s_x als Standardabweichung der im Test erzielten Rohwerte und r dem Reliabilitätskoeffizienten des Testes.

Im Grundintelligenztest CFT3 Skala 3 beträgt für die Rohwerte des Gesamttestes die Standardabweichung $s_x = 10,6$. Bei einem Reliabilitätskoeffizienten von $r_{tt} = 0,91$ ergibt sich ein Standardmeßfehler von

$$s_e = s_x \sqrt{1-r} = 10,6 \sqrt{1-0,91} = 3,18$$

Rohwerten³³.

In 68 % der Fälle wird der wahre Rohwert innerhalb des Bereichs $\pm 3,18$ liegen (vgl. die Normalverteilung). $x \pm 1 \cdot s_e$ ist das „Vertrauensintervall“ eines Rohwertes x , bezeichnet mit CL_x (Confidence Limits). Der wahre Rohwert eines Probanden A, der von maximal 100 Rohwerten 40 Rohwerte erzielt hat, liegt im Bereich von $40 \pm 3,18$ Rohwerte, also eine Standardmeßfehlergröße über und unter seinem erzielten Rohwert.

In 95 % der Fälle wird der wahre Rohwert im Bereich von zwei Standardmeßfehlergrößen über oder unter dem erzielten Rohwert liegen: $x \pm 2 \cdot s_e$ gibt das Vertrauensintervall eines Rohwertes mit 95 %iger Sicherheit an. Statt von 95 %iger Sicherheit wird auch oft von 5 %iger Irrtumswahrscheinlichkeit gesprochen [$100 \% - 95 \% = 5 \% = p$ (Probability)]. Dies ist eine gleichwertige Aussage. Der Rohwert des Probanden A würde mit 95 %iger statistischer Sicherheit im Bereich von $40 \pm 2 \cdot 3,18 = 40 \pm 6,36 \approx 40 \pm 6$ Rohwerten liegen. Diese Vertrauensbereichsgrenzen führen in der Normtabelle des CFT 3 zu dem entsprechenden IQ-Bereich von IQ = 89 bis IQ = 105.

Definition 26:

Das *Vertrauensintervall* CL_x (Confidence Limits) gibt den Bereich an, in dem der wahre Rohwert eines Probanden liegt.

$CL_x = x \pm 1 \cdot s_e$ (68 %ige statistische Sicherheit),

$CL_x = x \pm 2 \cdot s_e$ (95 %ige statistische Sicherheit),

mit

CL_x Vertrauensbereich in Rohwerten,

x Rohwert,

s_e Standardmeßfehler.

Im o.g. Beispiel wurde der Standardmeßfehler über die Rohwerte berechnet: verwandt wurde die Standardabweichung der Rohwerte. Dies führte zum Standardmeßfehler in Rohwerten, der dann einen Vertrauensbereich in Rohwerten festlegte, sei es mit 68- oder 95 %iger Sicherheit. Aus der Normtabelle ließ sich dann das entsprechende Vertrauensintervall in IQ-Einheiten ablesen, indem zum größt- und zum kleinstmöglichen Rohwert die IQ Werte aufgesucht wurden.

Das Vertrauensintervall kann nun aber auch direkt für die IQ-Skala oder eine andere Skala berechnet werden, ohne den Umweg über die Rohwerte zu gehen. Dazu wird der Standardmeßfehler nicht wie oben in Rohwerte sondern gleich in der Skalaeinheit ausgedrückt, die in den Normen des Testes verwendet wird. Für die Drahtbiegeprobe ist so vorgefahren worden: (1) Für die gewählte Skala $Z = 100 + 10 \cdot z$ (Standardwertskala „SW“) beträgt die Standardabweichung „groß“ ≈ 10 Einheiten.

In der Formel $s_e = s_x \sqrt{1-r}$ wird statt der Standardabweichung s_x der Rohwerte die Standardabweichung S der Skala eingesetzt. $s_e = S \sqrt{1-r}$. Unter Zugrundelegung eines Reliabilitätskoeffizienten von $r = 0,86$ (Mittelwert aller Einzelberechnungen der Reliabilität³⁴) ergibt sich für die Drahtbiegeprobe ein Standardmeßfehler von

$$s_e = S \sqrt{1-r} = 10 \cdot \sqrt{1-0,86} = 3,74.$$

Der Vertrauensbereich CL_z in der Skala Z beträgt dann mit 95 %iger statistischer Sicherheit:

$$CL_z = SW \pm 2 \cdot s_e = SW \pm 2 \cdot 3,74 = SW \pm 7,48.$$

Erreicht ein Proband in der Drahtbiegeprobe einen Standardwert von $SW = 105$, dann beträgt bei 95 %iger Sicherheit der Vertrauensbereich dieser Leistung $CL_z = 105 \pm 7,48$.

Die Bedeutsamkeit interindividueller Unterschiede in einem Test

Bei einem Leistungsvergleich zweier Probanden erhebt sich die Frage, ab welcher Leistungsdifferenz ein statistisch signifikanter, d. h. ein vom Zufall bzw. vom Meßfehler unabhängiger Unterschied vorliegt. Diese Frage läßt sich über die Vertrauensintervalle beantworten. Überschneiden sich die Intervalle zweier Probanden überhaupt nicht, so kann mit Sicherheit gesagt werden, daß zwischen beiden ein Leistungsunterschied besteht; der Leistungsunterschied ist auf jeden Fall statistisch signifikant.

In einem vom Lehrer selbst erstellten Test mit z. B. maximal 40 erreichbaren Rohwerten möge der Standardmeßfehler $s_e = 1,2$ Rohwerte betragen. Der Vertrauensbereich wird $2 \cdot s_e = 2,4$ Rohwerte über und unter einem gegebenen Rohwert liegen. (95 %ige Sicherheit.) Erreichte Schüler A in diesem Test 12 Rohwerte und Schüler B 18, unterscheiden sich beide Schüler statistisch signifikant. Die Vertrauensintervalle überschneiden sich nicht (Abbildung 25).



Abb. 25:

Vertrauensintervalle bei $s_e = 1,2$ und 95 %iger Sicherheit ($CL_x = x \pm 2 \cdot 1,2$).

Wollen Sie einen sehr strengen prüfstatistischen Maßstab anlegen, können Sie von vornherein einen statistisch signifikanten Leistungsunterschied ausschließen, wenn sich die Vertrauensintervalle überschneiden. Im allgemeinen aber wird bei einem Überschneiden der Vertrauensintervalle nach dem Grad der Überschneidung gefragt.

Die sogenannte kritische Differenz bestimmt, bei welchem Grad der Überschneidung der Vertrauensintervalle noch von einem statistisch signifikanten Leistungsunterschied gesprochen werden kann. Ohne Ableitung sei hier die Formel für die kritische Differenz angegeben³⁵. Sie lautet bei 95 %iger statistischer Sicherheit:

$$(x_1 - x_2)_{0,05} = 2 \cdot s_x \sqrt{2(1 - r_{tt})}.$$

Dies gibt die kritische Differenz in Rohwerten an. Die Zahl 0,05 im Index bei $(x_1 - x_2)$ soll verdeutlichen, daß die rechte Seite der Gleichung für eine Irrtumswahrscheinlichkeit von 5 % gilt, d. h. für eine statistische Sicherheit von 95 %. Soll die kritische Differenz gleich in der Standardwert-Skala angegeben werden, die ein Test verwendet, muß statt s_x die Standardabweichung „groß“ S der betreffenden Skala eingesetzt werden. Beträgt bei einem Intelligenztest der Reliabilitätskoeffizient $r = 0,85$ und wird die IQ-Skala $100 + 15 \cdot z$ benutzt, dann beträgt die kritische Differenz in IQ-Einheiten bei 95 %iger Sicherheit:

$$\begin{aligned} (IQ_1 - IQ_2)_{0,05} &= 2 \cdot S \sqrt{2(1 - r_{tt})} \\ &= 2 \cdot 15 \sqrt{2(1 - 0,85)} \\ &= 16,43. \end{aligned}$$

Unterscheiden sich zwei Probanden in ihrer Intelligenzleistung um 16,43 IQ Einheiten, so kann mit hoher Wahrscheinlichkeit ($p = 0,05$) gesagt werden, daß die Intelligenz der zwei Probanden, wie sie von dem betreffenden Test gemessen wird, verschieden ist.

Definition 27:

Über die *kritische Differenz* wird bestimmt, bei welchem Leistungsunterschied zwischen zwei Probanden in einem Test noch von einem – im prüfstatistischen Sinne – signifikanten Unterschied gesprochen werden kann. Bei 95 %iger statistischer Sicherheit ($p = 0,05$) bestimmt sich die kritische Differenz in Rohwerten über:

$$(x_1 - x_2)_{0,05} = 2 \cdot s_x \sqrt{2(1 - r_{tt})}$$

mit

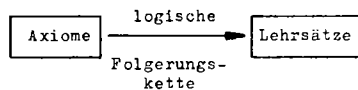
- | | |
|------------|--|
| x_1, x_2 | Rohwertleistung des Probanden 1
bzw. des Probanden 2, |
| s_x | Standardabweichung der Rohwerte des Testes, |
| r_{tt} | Reliabilitätskoeffizient des Testes. |

Liegt im prüfstatistischen Sinne ein signifikanter Unterschied zwischen zwei Probanden vor, so ist aber noch festzustellen, ob dieser Unterschied auch bedeutsam ist. Wenn von einem wirklichen Unterschied gesprochen werden soll, muß eine Leistungsdifferenz zwischen 12 und 18 Rohwerten wie in Abbildung 25 aus pädagogischen oder psychologischen Gründen auch als bedeutsam angesehen werden können, bezogen auf diesen Test.

9.6 Exkurs: Die Grundaussagen (Axiome) der klassischen Testtheorie

Die klassische Testtheorie, wie sie hier dem Kapitel über die Reliabilität zugrunde liegt, läßt sich auf bestimmte Grundaussagen (Axiome) zurückführen. Die Ableitung des Standardmeßfehlers geht z. B. von diesen Axiomen aus. Im folgenden soll auf diese Axiome eingegangen werden. Vorangestellt seien einige Bemerkungen über das Wesen eines Axioms.

In der Mathematik lassen sich Aussagen auf möglichst einfache Grundaussagen – sogenannte Axiome – zurückführen. Diese Axiome sind zu-
meist von solch einfachem Inhalt, daß vor allem der mathematische Laie weder das Verlangen noch die Notwendigkeit für einen Beweis verspürt (z. B. aus der Axiomatik der reellen Algebra: $a = b$ daraus folgt $b = a$). Axiome sind brauchbar, wenn sich aus ihnen keine widersprüchlichen Aussagen herleiten lassen. Die Aussagen, die aus den Axiomen über logische Schlußfolgerungen abgeleitet werden, sind die Lehrsätze.



Ähnlich ist es nun in der klassischen Testtheorie. Die Formel für den Standardmeßfehler wird aus den Axiomen der klassischen Testtheorie logisch abgeleitet. Die „Wahrheit“ dieser Formel steht und fällt mit der Frage, ob die Axiome der klassischen Testtheorie „wahr“ sind. Diese Axiome sind:

1. *Existenzaxiom*: Jeder aktuelle (gemessene) Wert setzt sich aus einem wahren Wert und einem Fehlerwert zusammen. Der wahre Wert wird als eine Konstante angenommen, zumindest über einen gewissen Zeitraum hinweg.
2. *Meßfehleraxiom*: Die Meßfehler (Differenzen zwischen aktuellen Werten und wahren Wert) bei einem Probanden sind bei wiederholten Messungen teils positiv, teils negativ. Sie verteilen sich auf lange Sicht so, daß ihr Mittelwert gleich Null ist. M. a. W. der Meßfehler einer Messung ist eine Zufallsvariable, wobei sein Mittelwert Null beträgt. Wahre Werte und Meßfehler sind unkorreliert.

3. *Verknüpfungsbaxiom*: Wahre Werte und Meßfehler sind additiv miteinander verknüpft. So wird bei der Ableitung des Standardmeßfehlers davon ausgegangen, daß die aktuelle Varianz der Testwerte (s_x^2) sich aus der Summe der wahren Varianz (s_w^2) und der Meßfehlervarianz (s_e^2) zusammensetzt ($s_x^2 = s_w^2 + s_e^2$).

Die Kritik an der Klassischen Testtheorie setzt einmal an dem 2. Axiom an. Wenn der Meßfehler einer Messung eine Zufallsvariable ist, heißt dies, daß die Meßfehler völlig unabhängig von den wahren Werten auftreten. Können aber nicht unterschiedliche wahre Werte auch unterschiedliche Meßfehlerwerte hervorrufen? Sind die Meßfehlerwerte nicht nur von der mangelnden Reliabilität des Testes sondern auch von der jeweils untersuchten Population abhängig? (z. B. könnte der Meßfehler bei Leistungstarken eine andere Größe einnehmen als bei Leistungsschwachen). Trifft dies zu, dürften sich wahre Werte und Meßfehlerwerte nicht mehr einfach additiv verknüpfen lassen. Die Ableitung des Standardmeßfehlers wäre hinfällig.

Die klassische Testtheorie ist an der Normalverteilung ausgerichtet. Soll in einem Schulleistungstest lediglich festgestellt werden, ob die Schüler ein Lernziel erreicht haben und interessiert dabei nicht, wie gut oder wie schlecht das Lernziel erreicht worden ist (kriterienbezogene Leistungsmessung), dann versagt die klassische Testtheorie. Varianzen bzw. Standardabweichungen lassen sich nur für annähernd normal verteilte Leistungsverteilungen berechnen: Bei kriterienbezogenen Tests gibt es in dem Sinne keine normalen Verteilungen.

In der psychologischen Meßtheorie werden neue Modellansätze entwickelt, eingeleitet durch das Rasch-Modell (1960). Die klassische Testtheorie wird dadurch nicht verworfen werden. Für Tests im Unterricht hat die normbezogene Leistungsmessung, fußend auf einer klassischen Testtheorie, und die kriterienbezogene Leistungsmessung, fußend auf einer neuen Testtheorie, ihren jeweils eigenen pädagogischen Stellenwert³⁶. – Die Entwicklung bei den psychologischen Meßverfahren, wie z. B. bei den Intelligenztests, ist nicht abzusehen. Festgehalten werden kann, daß die derzeit auf dem Markt befindlichen psychologischen Tests sich in der Regel alle auf der klassischen Testtheorie gründen.

10. Objektivität

10.1 Der Begriff der Testobjektivität

Die Objektivität eines Testes bezieht sich darauf, zu welchem Grad die Ergebnisse eines Testes unabhängig vom Untersucher sind. Ein Test muß unabhängig von der Person sein, die ihn als Testleiter durchführt, die ihn auswertet und die die Leistungen interpretiert. Man spricht demzufolge von der Durchführungs-, Auswertungs- und Interpretationsobjektivität. Es wird unten auf diese Aspekte der Testobjektivität noch eingegangen werden.

Im allgemeinen Sprachgebrauch wird ein Testresultat dann als „objektiv“ bezeichnet, wenn es reliabel und valide ist. So gesehen könnte man, wenn man es mit reliablen Testwerten zu tun hat, auch von objektiven Testwerten sprechen. In der Testtheorie jedoch wird unterschieden zwischen der Objektivität, Reliabilität und Validität eines Testes, die auch als die Gütekriterien eines Testes bezeichnet werden. Der Begriff der Objektivität wird auf die o.g. Bedeutung einer Durchführungs-, Auswertungs- und Interpretationsobjektivität eingegrenzt.

Die Durchführungs- und Auswertungsobjektivität sind eine notwendige Voraussetzung für die Reliabilität eines Testes. Variieren die Testergebnisse mit der Person, die einen Test mit Probanden durchführt, und variieren sie mit der Person, die ihn auswertet, kann der Test nicht zu genauen, m.a.W. zu reliablen Ergebnissen führen. Auf der anderen Seite ist eine gegebene Objektivität bei der Durchführung und Auswertung keine hinreichende Bedingung für die Reliabilität. Ist nämlich ein Test so angelegt, daß er zu objektiven (d. h. vom Untersucher unabhängigen) Werten führt, so müssen diese Werte nicht notwendigerweise genau (reliabel) ermittelt sein.

Die Interpretationsobjektivität ist eine notwendige Bedingung für die Validität eines Testes. Sind die Aussagen, die anhand der Testergebnisse gemacht werden, abhängig von der Person, welche die Testergebnisse interpretiert, so kann der Test nicht zu gültigen, m.a.W. zu validen Aussagen führen. Andererseits ist aber die Interpretationsobjektivität keine hinreichende Bedingung für die Validität eines Testes. Testergebnisse, deren Interpretationsobjektivität gesichert ist, müssen nicht notwendigerweise zu validen Aussagen führen; d. h., eine objektiv nach den Testanweisungen abgeleitete Interpretation muß nicht notwendigerweise eine gültige Aussage über die zu messende Fähigkeit machen.

Bereits in den Abhandlungen zur Reliabilität wurde herausgestellt, daß die Reliabilität eine notwendige aber nicht hinreichende Bedingung für die Validität ist. Die dort in Abbildung 22 aufgestellte Beziehung zwi-

schen den zwei Gütekriterien Validität und Reliabilität läßt sich nun um das Gütekriterium der Objektivität erweitern (Abbildung 26).

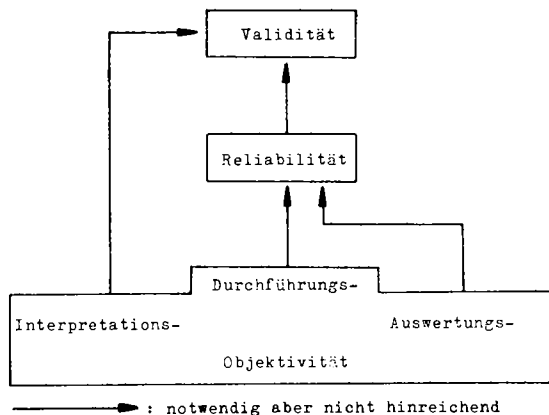


Abb. 26:

Beziehung zwischen der Objektivität, Reliabilität und Validität eines Testes.

10.2 Durchführungsobjektivität

Die Durchführungsobjektivität ist gegeben, wenn die äußeren Bedingungen der Testsituation für alle Testpersonen gleich sind und die Anweisungen des Testautors im Detail schriftlich so festgehalten sind, daß eine unterschiedliche Auslegung der Durchführungsanweisungen nicht mehr möglich ist. Dem Testleiter wird in der Handanweisung mitgeteilt, was er bei der Durchführung tun und sagen muß, welche Hilfen oder Hilfsmittel erlaubt oder verboten sind, etc. Den Anweisungen muß exakt gefolgt werden. Man spricht von den „standardisierten“ Bedingungen der Testdurchführung.

Bei vom Lehrer konstruierten Schulleistungstests muß z. B. festgelegt werden, welche Bearbeitungszeit gegeben wird, welche Hilfsmittel die Schüler benutzen können (z. B. Taschenrechner bei einem Mathematiktest), welche Hilfen der Lehrer geben kann. – Das Problem der Durchführungsobjektivität wird sich bei einem Schulleistungstest weniger stellen. Problematisch kann hier die Auswertungsobjektivität werden.

10.3 Auswertungsobjektivität

Die Auswertungsobjektivität ist gegeben, wenn das Auswertungsergebnis bei einem Test unabhängig davon ist, welche Person den Test auswertet.

Dazu ist es notwendig, daß vom Testautor genau festgelegt ist, was als richtig oder als falsch zu bewerten ist.

Der Grad der Auswertungsobjektivität ist u. a. abhängig von den Aufgabenformen, die in einem Test verwandt sind. Bei Richtig-Falsch Aufgaben oder Mehrfachwahlaufgaben mit einer eindeutigen Lösung ist die Auswertungsobjektivität praktisch vollkommen verwirklicht. Die Auswertung kann mit Schablonen durchgeführt werden.

Das Paradebeispiel für eine eingeschränkte Auswertungsobjektivität ist der Schulaufsatz. Der subjektive Ermessensspielraum des Testleiters ist bei der „Auswertung“ eines Aufsatzes groß. Dies kann auch auf Testaufgaben zutreffen, die nur eine Kurzantwort verlangen. Im eingangs vorgestellten Beispielttest Vitamine und Mineralstoffe sind eine Reihe von Kurzantwortaufgaben aufgenommen. Wenn Sie Ihre Antworten zu diesen Aufgaben mit denen der Musterlösung vergleichen, werden Sie ermessen, welche Schwierigkeiten die Auswertungen von Kurzantwortaufgaben machen können. Besondere Probleme werden bei diesem Test die Mehrfachwahlantwortaufgaben machen. Es sind hier Mehrfachantworten zugelassen. Bei Aufgabe 8 z. B. gelten vier Antworten aus sechs Antwortvorgaben als richtig. Wenn nur drei Antworten einer Testperson richtig sind, wird diese Aufgabe dann noch als gelöst angesehen und mit einem Punkt bewertet? Reicht es noch aus, wenn aus den sechs Vorgaben zwei als richtig erkannt wurden? Der Testentwickler muß präzise Auswertungsanweisungen von Aufgabe zu Aufgabe angeben, will er sicherstellen, daß auch ein anderer Auswerter neben ihm selbst zu dem gleichen Auswertungsergebnis kommt.

10.4 Interpretationsobjektivität

Die Interpretationsobjektivität liegt vor, wenn zwei Beurteiler unabhängig voneinander aus dem gleichen Auswertungsergebnis den gleichen Schluß ziehen. M.a.W. das gleiche Auswertungsergebnis muß von zwei Beurteilern unabhängig voneinander in gleicher Weise interpretiert werden.

Handelt es sich um standardisierte Intelligenztests, kann die Interpretationsobjektivität als trivial angesehen werden. Wenn in einem Intelligenztest die Rohwertleistung eines Probanden sich anhand der Normen in einen IQ-Wert umrechnen läßt, ist die Interpretation des Auswertungsergebnisses festgelegt. Der Vergleichsmaßstab, die Norm des Testes, ist für alle Beurteiler gleich. Die Interpretationsobjektivität kann hier als künstlicher Aspekt angesehen werden; die Interpretation des Testergebnisses fällt unter die Auswertung.

Offensichtlicher wird der Unterschied zwischen der Auswertungs- und Interpretationsobjektivität bei nichtstandardisierten Tests. Bei einem informellen Leistungstest wie dem Test über Vitamine und Mineralstoffe wird jeder Lehrer seinen eigenen Maßstab zur Beurteilung des Auswertungsergebnisses anlegen, vorausgesetzt, das Auswertungsergebnis läßt sich objektiv ermitteln. Ein Lehrer in einer leistungsstarken Klasse würde u. U. eine Leistung von acht aus 20 möglichen Rohwerten als „mangelhafte“ Leistung ansehen. Ein anderer könnte diese Leistung noch durchaus als „ausreichend“ interpretieren, wenn er von einer insgesamt leistungsschwachen Klasse ausgeht. Von unterschiedlichen Normen ausgehend, gelangen diese Lehrer zu verschiedenen Interpretationen. Weder die leistungsschwache noch die leistungsstarke Klasse ist repräsentativ. Erst wenn mehrere Klassen diesem (vorher mehrfach revidierten) Test unterzogen werden und aus den Ergebnissen aller Klassen eine Normentabelle angelegt wird, gewinnt der Test an Interpretationsobjektivität.

11. Aufgabenanalyse

Nach der ersten Erprobung eines Testes müssen die Testaufgaben eingehend analysiert werden. Es müssen die Aufgaben aufgespürt werden, die wenig zur Reliabilität und Validität des gesamten Testes beitragen. Dies leistet eine Aufgabenanalyse, auch Item-Analyse genannt. Sie verfolgt drei Ziele:

- die Bestimmung des *Schwierigkeitsgrades* der einzelnen Aufgaben,
- die Ermittlung der *Trennschärfe* der einzelnen Aufgaben (Feststellung, wie gut eine Aufgabe in der Lage ist, zwischen leistungsstarken und leistungsschwachen Personen zu differenzieren),
- und bei Mehrfachwahlaufgaben die *Distraktorenanalyse* (Untersuchung, ob die falschen Alternativen einer Aufgabe eher von den im Gesamttest leistungsschwachen Personen für die richtige Antwort gehalten werden als von den leistungsstarken).

11.1 Schwierigkeitsgrad und Trennschärfe

Schwierigkeitsgrad

Der Schwierigkeitsgrad einer Aufgabe gibt an, zu welchem Anteil die Getesteten die Aufgabe richtig lösten. Ein Schwierigkeitsgrad bei einer Aufgabe von $p = 0,20$ gibt an, daß nur 20 % der Testpersonen diese Aufgabe lösten. Die Aufgabe ist schwer: 80 % konnten sie nicht lösen. Bei einem p von 0,85 ist eine Aufgabe leicht. 85 % lösten die Aufgabe; nur 15 % konnten sie nicht lösen. Ist der numerische Wert von p hoch, ist die Aufgabe leicht. Wenn der p -Wert niedrig ist, gilt die Aufgabe als schwer (s. Abbildung 27).

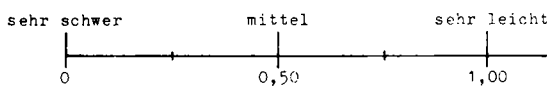


Abb. 27:

Die Reichweite des Schwierigkeitsgrades p

Definition 28:

Der *Schwierigkeitsgrad* einer Aufgabe gibt an, zu welchem Anteil eine Personengruppe die Aufgabe richtig löste.

$$p = \frac{N_r}{N}$$

mit

p = Schwierigkeitsgrad einer Aufgabe,

N_r = Anzahl der Personen, die die Aufgabe,
richtig lösten,

N = Gesamtzahl der Personen.

Aufgaben, die von nahezu allen oder von nahezu niemandem gelöst werden, tragen zur Differenzierung zwischen den leistungsstarken und leistungsschwachen Personen nichts bei. In einem möglichst hoch differenzierenden Test müssen die extrem leichten und extrem schweren Aufgaben eliminiert werden. Ob eine Aufgabe eliminiert, verändert oder beibehalten werden soll, ist allerdings aus dem Schwierigkeitsgrad der Aufgabe allein nicht abzulesen. Entscheidend ist die Trennschärfe der Aufgabe.

Trennschärfe

Mit der Trennschärfenbestimmung einer Aufgabe soll angegeben werden, wie gut eine Aufgabe in der Lage ist, zwischen leistungsstarken und leistungsschwachen Personen in einem Test zu unterscheiden.

Der Trennschärfenbestimmung liegt folgender Gedanke zugrunde. Jede Aufgabe eines Testes kann als ein Test in sich aufgefaßt werden. Die einzelne Testaufgabe muß das Gleiche messen, wie alle übrigen Testaufgaben zusammen. Über einen Korrelationskoeffizienten läßt sich bestimmen, inwieweit eine Testaufgabe auch das mißt, was die restlichen Testaufgaben zusammen messen. Abbildung 28 verdeutlicht das Vorgehen. Der große Kreis unter A steht für einen informellen Schulleistungstest mit 30 Aufgaben. Aufgabe 14 wird herausgezogen. Die Ergebnisse im „Test“ 14 werden mit den Ergebnissen im gesamten Test ohne die Aufgabe 14 korreliert (B). Dies ergibt den Trennschärfekoeffizienten r_{it} für die Aufgabe 14. Die Buchstaben i im Index von r stehen für „Item Test“ da ja das Item (die Aufgabe) mit dem gesamten Test korreliert wird. Soll der Trennschärfekoeffizient von Aufgabe 15 berechnet werden, wird

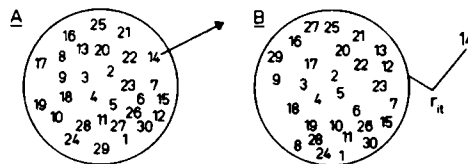


Abb. 28:

Bestimmung des Trennschärfekoeffizienten: Korrelation Item-Test.

Aufgabe 14 wieder dem gesamten Test zugeordnet, Aufgabe 15 herausgezogen und die Korrelation zwischen dieser Aufgabe und dem gesamten Test ohne diese Aufgabe berechnet.

Eine Datentabelle für die Berechnung des Trennschärfekoeffizienten für Aufgabe 14 wird so aussehen, wie sie in Tabelle 28 angelegt ist.

Dieser Proband erreichte im Test ohne Aufgabe 14 4 Rohwerte; A.14 konnte er nicht lösen.

Aufgabe 14	Anzahl der erzielten Rohwerte im gesamten Test ohne Aufgabe 14																												
	0	1	2	3	4	5	6	7	8		25	26	27	28	29														
0 (ungelöst)					①																								
1 (gelöst)																													

Tab. 28:

Linkes und rechtes Ende eines Auswertungsblattes (Strichliste) zur Berechnung eines Trennschärfekoeffizienten.

Der gesamte Test enthält 30 Aufgaben. Pro Aufgabe wird ein Rohwert vergeben. Maximal sind 30 Rohwerte zu erreichen. Wird eine Aufgabe herausgezogen, sind maximal nur noch 29 Rohwerte zu erreichen. Für die Gesamtrohwertzahl sind in der Waagerechten, am Kopf der Tabelle 28, 29 Einteilungen vorgenommen. War Aufgabe 14 richtig gelöst, wird 1 Rohwert, war sie falsch gelöst, wird kein Rohwert (0) vergeben. Für Aufgabe 14 sind in der Senkrechten am linken Ende der Tabelle 28 zwei Einteilungen (0/1) vorgenommen. Jeder Proband wird über einen Strich in der Tabelle vermerkt, gemäß seiner Gesamtrohwertzahl und seiner Leistung bei Aufgabe 14.

Die Aufgabe 14 ist nun trennscharf, wenn sie von den leistungsstarken Schülern häufiger gelöst wird als von den leistungsschwachen. Dies trifft für die Eintragungen in Tabelle 28 zu. Die Schüler, die insgesamt auch am besten abgeschnitten haben, also die Schüler mit insgesamt 26, 27, 28 oder sogar 29 Rohwerten, lösten in der Regel auch Aufgabe 14. Nur einer von ihnen löste die Aufgabe nicht. Jene Schüler, die insgesamt schlechter abgeschnitten haben, also die Schüler mit nur insgesamt 4, 5, 6, 7 oder 8 Rohwerten, beantworteten Aufgabe 14 in der Regel auch falsch. Nur drei von ihnen konnten die Aufgabe lösen.

Häufen sich die Auszählungsstriche in Tabelle 28 rechts unten und links oben, so ist eine Aufgabe trennscharf. Eine Aufgabe ist nicht trennscharf,

wenn sowohl einige der leistungsstarken als auch einige der leistungsschwachen Schüler die Aufgabe falsch bzw. richtig gelöst haben. In Tabelle 28 würden sich die Auszählungsstriche gleichmäßig verteilen. Eine Aufgabe wird negativ trennscharf, wenn die leistungsschwachen Schüler zu hohen Anteilen eine Aufgabe richtig lösten, die leistungsstarken aber die Aufgabe in der Regel falsch lösten. In Tabelle 28 würden sich die Auszählungsstriche dann rechts oben und links unten häufen. Der errechnete Korrelationskoeffizient (Trennschärfekoeffizient) würde einen negativen Wert annehmen.

Die zwei Variablen in Tabelle 28 (Aufgabe und Testrohwert) haben unterschiedliches Skalenniveau. Das Skalenniveau der Aufgabe ist alternativ (0/1), das des Testrohwertes quantitativ (0, 1, 2, ... 29) verteilt. Soll hier der Korrelationskoeffizient berechnet werden – dies ist ja der Trennschärfekoeffizient – muß die sog. punktbiseriale Korrelation r_{pbis} berechnet werden³⁷. (Im Kapitel über Korrelation wurde nur der Pearson-Produkt-Moment Korrelationskoeffizient vorgestellt, der angewandt wird, wenn zwei Variablen mit quantitativem Skalenniveau miteinander zu korrelieren sind.) Der punktbiseriale Korrelationskoeffizient ist aus dem Pearson-Produkt-Moment Korrelationskoeffizienten abgeleitet.

Definition 29:

Die *Trennschärfe* einer Aufgabe gibt an, wie gut die Aufgabe zwischen den leistungsstarken und leistungsschwachen Personen in einem Test unterscheiden kann. Die Trennschärfe wird angegeben über einen Trennschärfekoeffizienten. Dies ist die punktbiseriale Korrelation r_{pbis} zwischen den Ergebnissen einer Aufgabe und den Ergebnissen im gesamten Test ohne diese Aufgabe.

Es ist für die Konstruktion eines informellen Schulleistungstests ein aufwendiges Verfahren, wenn für jede Aufgabe eines Testes der punktbiseriale Korrelationskoeffizient berechnet werden muß; aus diesem Grunde wird seine Formel hier nicht angegeben. Im folgenden wird ein vereinfachtes Verfahren zur Bestimmung der Trennschärfe einer Aufgabe vorgestellt. Über dieses Verfahren lassen sich die Trennschärfekoeffizienten, die sonst über die Korrelation ermittelt werden, relativ genau schätzen. Neben der Trennschärfe läßt sich über dieses Verfahren auch der Schwierigkeitsgrad bestimmen.

11.2 Vereinfachtes Verfahren zur Trennschärfe- und Schwierigkeitsbestimmung (Diskriminationsindex D)

Die Bestimmung der Aufgabentrennschärfe und Aufgabenschwierigkeit nach einem vereinfachten Verfahren (Diskriminationsindex oder D-index for item analysis) gliedert sich in sechs Schritte³⁸:

1. Die Rohwerte eines Testes werden in eine Rangreihe gebracht.
2. Die Gesamtzahl N der Getesteten wird mit 0,27 multipliziert und das Ergebnis zur nächstgelegenen ganzen Zahl auf- bzw. abgerundet. Diese Zahl wird n genannt. Wenn N = 30 ist, würde n = 8 sein (8,1 abgerundet).
3. Die in 2 ermittelte Zahl n bestimmt den Umfang einer „hohen“ Gruppe (die n Leistungsstärksten) und einer „niedrigen“ Gruppe (die n Leistungsschwächsten).
4. In der hohen und der niedrigen Gruppe wird für jede Aufgabe der Anteil der richtigen Antworten bestimmt:

$$p_{\text{Hoch}} = \frac{\text{Zahl der richtigen Antworten in der "hohen" Gruppe}}{n}$$

$$p_{\text{Niedrig}} = \frac{\text{Zahl der richtigen Antworten in der "niedrigen" Gruppe}}{n}$$

5. Der Schwierigkeitsgrad einer Aufgabe gibt den Anteil der Testpersonen an, der die Aufgabe richtig gelöst hat. Aus den Werten p_{Hoch} und p_{Niedrig} läßt sich der Schwierigkeitsgrad p in guter Näherung über das arithmetische Mittel aus beiden Größen angeben:

$$p = \frac{p_{\text{Hoch}} + p_{\text{Niedrig}}}{2}$$

6. Um den Trennschärfekoeffizienten für eine Aufgabe in guter Näherung zu erhalten, wird p Niedrig von p Hoch subtrahiert. Dies ergibt den Trennschärfenindex D (Diskriminationsindex):

$$D = p_{\text{Hoch}} - p_{\text{Niedrig}}.$$

Tabelle 29 zeigt, welche Werte der Diskriminationsindex, d. h. also die Trennschärfe, und der Schwierigkeitsgrad einer Aufgabe annehmen können. Nehmen Sie an, ein informeller Schulleistungstest wurde von N = 30 Schülern bearbeitet. Hier wäre $n = 0,27 \cdot N = 0,27 \cdot 30 = 8,1$. Die Testbögen der im Gesamttest acht besten Schüler kommen in die „hohe“ Gruppe, und die der 8 schlechtesten in die „niedrige“ Gruppe.

Aufgabe 1 lösten 7 Schüler der hohen Gruppe richtig, während nur 2 der niedrigen Gruppe die richtige Lösung fanden. Vom meßtheoretischen Standpunkt aus arbeitet Aufgabe 1 richtig. Sie trägt zur Leistungsdifferenzierung bei. Die leistungsstarken Schüler haben diese Aufgabe häufi-

ger richtig gelöst als die leistungsschwachen. Die Trennschärfe ist mit 0,63 hoch. Der Schwierigkeitsgrad liegt mit 0,57 günstig (mittlerer Bereich).

Aufgabe 2 ist mit $p = 0,94$ zu leicht, um noch eine ausreichende Trennschärfe gewährleisten zu können ($D = 0,12$).

Aufgabe 3 liegt mit $p = 0,50$ im mittleren Schwierigkeitsbereich, erweist sich jedoch nicht als trennscharf. Die leistungsstarken Schüler haben zu gleichen Anteilen wie die leistungsschwachen diese Aufgabe richtig bzw. falsch gelöst.

Aufgabe 3 zeigt, daß das Vorliegen eines annehmbaren Schwierigkeitsbereiches noch nicht bedeuten muß, daß eine Aufgabe auch trennscharf ist. Daß der Schwierigkeitsgrad einer Aufgabe im mittleren Bereich von $p = 0,35$ bis $p = 0,75$ liegt, ist nur eine notwendige, nicht aber eine hinreichende Bedingung für die Trennschärfe einer Testaufgabe.

Aufgabe	Anzahl der richtigen Antworten je Gruppe	Relativer Anteil der richtigen Antworten	Trennschärfe: Diskriminationsindex $D = P_H - P_N$	Schwierigkeitsgrad $p = \frac{P_H + P_N}{2}$
1	H 7 N 2	$7/8 = 0,88$ $2/8 = 0,25$	$0,88 - 0,25 = 0,63$	$\frac{0,88 + 0,25}{2} = 0,57$
2	H 8 N 7	1,00 0,88	0,12	0,94
3	H 4 N 4	0,50 0,50	0,00	0,50
4	H 3 N 7	0,38 0,88	-0,50	0,63

Tab. 29:

Aufgabenanalysedaten nach dem Diskriminations-Index-Verfahren für 4 Aufgaben, $N = 30$, H: „Hohe Gruppe“, N: „Niedrige Gruppe“ (nach J. C. Stanley, K. D. Hopkins, Educational and Psychological Measurement and Evaluation, Englewood Cliffs, New Jersey 1972, S. 274).

Aufgabe 4 ist negativ trennscharf. Die leistungsstarken Schüler lösten die Aufgabe falsch, während gerade die leistungsschwachen sie richtig lösten.

Aufgabe 1 und 2 können für eine veränderte Testfassung beibehalten werden. Die Testaufgabe 1 ist ein ausgezeichnetes „Item“. Aufgabe 2 ist zwar keine trennscharfe, aber eine leichte Aufgabe, die zum Testeinstieg gestellt werden kann: am Anfang des gesamten Testes sollten zwei bis drei leichte Aufgaben stehen. Die Aufgaben 3 und 4 müssen revidiert werden.

Ein Gespräch über diese Aufgaben mit den Schülern wird dem Testkonstrukteur in der Regel schnell zeigen, warum gerade diese Aufgaben nicht ihre Funktion erfüllen.

Bei Mehrfachwahlaufgaben läßt sich feststellen, welche Alternativen (Distraktoren) Schwierigkeiten bereiten. Für diese Testaufgaben läßt sich die Aufgabenanalyse noch um die Distraktorenanalyse erweitern. Hier auf wird weiter unten noch eingegangen.

11.3 Zusammenhang zwischen Trennschärfe und Schwierigkeit: Aufgabenselektion

Aufgabe 2 in Tabelle 29 zeigt, daß eine sehr leichte Aufgabe nur eine geringe Trennschärfe aufweist: bei hohem p ist die Trennschärfe, d.h. der Diskriminationsindex niedrig. Das Entsprechende gilt für schwere Aufgaben: Bei niedrigem p ist auch der Trennschärfeindex niedrig. Aufgaben, die von allen Testpersonen richtig oder falsch beantwortet werden, besitzen keine Trennschärfe mehr; sie tragen nichts zur Differenzierung zwischen leistungsstarken und leistungsschwachen Personen bei. Die theoretische Beziehung zwischen Schwierigkeit und Trennschärfe verdeutlicht Abbildung 29 grafisch.

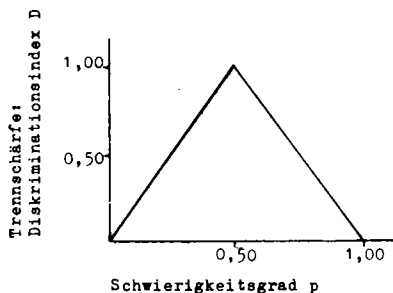


Abb. 29:

Theoretischer Zusammenhang zwischen Schwierigkeit und Trennschärfe

Erst bei mittlerer Aufgabenschwierigkeit (ca. $p = 0,5$) *kann* die Trennschärfe ihren maximalen Wert erreichen. Hier sei noch einmal darauf hingewiesen, daß die mittlere Schwierigkeit einer Testaufgabe nur eine notwendige, nicht aber eine hinreichende Bedingung für die Trennschärfe einer Aufgabe ist (vgl. Aufgabe 3 u. 4, Tabelle 29).

Von guten Aufgaben werden Trennschärfeindizes von 0,40 und höher erwartet (Tabelle 30).

Trennschärfe: Diskriminationsindex (D)	Interpretation
0,40 und höher	Ausgezeichnete Testaufgabe
0,30 bis 0,39	Brauchbare Testaufgabe, Verbesserung ist möglich
0,20 bis 0,29	Weniger brauchbare Test- aufgabe, Verbesserung ist notwendig
unter 0,19	Unbrauchbare Testaufgabe, Testausschluß oder Revision

Tab. 30:

Interpretation der Trennschärfeindizes (nach R. L. Ebel, Essentials of Educational Measurement, Englewood Cliffs, New Jersey 1972, S. 399).

Ein Diagramm erleichtert die Auswahl der brauchbaren Aufgaben (Abbildung 30). Aufgabe 17 z.B. in Abbildung 30 ist durch einen Schwierigkeitsgrad von $p = 0,55$ und durch einen Diskriminationsindex von $D = 0,63$ gekennzeichnet. Diese Aufgabe liegt innerhalb des Rechteckes, das die zu akzeptierenden Aufgaben umfaßt. In dieses Rechteck werden Aufgaben aufgenommen, die eine Schwierigkeit von $p = 0,20$ bis $p = 0,80$ und einen Trennschärfeindex von $D = 0,20$ bis $D = 1,00$ aufweisen. Aufgaben außerhalb dieses Rechteckes müssen gänzlich revidiert oder durch neue ersetzt werden. Aufgaben, die innerhalb des Rechteckes unter die gestrichelte Linie fallen, können zwar beibehalten werden, müssen jedoch verbessert werden. Ihr Trennschärfeindex D liegt nur zwischen 0,20 und 0,29.

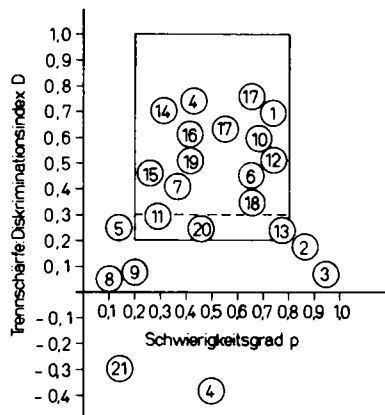


Abb. 30:

Hilfsdiagramm zur Selektion von Testaufgaben.

Die Selektion der Testaufgaben nach Trennschärfe und Schwierigkeitsgrad sollte nicht ausschließen, daß zwei oder drei leichte Aufgaben, die nicht in den Annahmehereich fallen, trotz alledem in den Test aufgenommen werden. Die Aufgaben 2 und 3 z. B. sind leicht und empfehlen sich, zum motivierenden Einstieg an den Testanfang gestellt zu werden. Danach folgen die Aufgaben, die im Annahmehereich des Testes liegen. Sofern nicht einer inhaltlichen Gruppierung der Testaufgaben der Vorrang gegeben wird, können die Aufgaben nach dem Schwierigkeitsgrad geordnet werden.

11.4 Distraktorenanalyse

Bei gut konstruierten Mehrfachwahlaufgaben muß zu erwarten sein, daß die falschen Alternativen, die Distraktoren, eher von den leistungsschwachen Schülern für die richtige Antwort gehalten werden als von den leistungsstarken.

Eine Mehrfachwahlaufgabe muß *extrinsisch*, darf aber nicht *intrinsisch* mehrdeutig angelegt sein. An einer Mehrfachwahlaufgabe sollen diese zwei Begriffe verdeutlicht werden.

Der Schüler soll den englischen Satz auswählen, der die beste Übersetzung des deutschen Satzes darstellt.

Das Grundstück gehört immer noch Vater.

- A Father always hears good pieces.
- B The piece of ground still belongs to Father.
- C The property still hears to Father.
- D The property still belongs to Father.

Die beste Übersetzung, Alternative D, ist mehr als nur eine wörtliche Übersetzung wie die in Alternative B. Antwort D trifft das Konzept, das

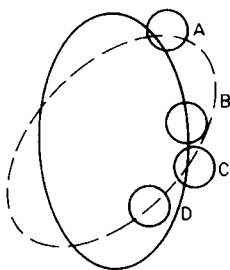


Abb. 31:

Verdeutlichung des semantischen Raumes einer Testaufgabe, die extrinsische Mehrdeutigkeit besitzt (nach J. C. Stanley, K. D. Hopkins, Educational and Psychological Measurement and Evaluation, Englewood Cliffs, New Jersey 1972, S. 234, modifiziert).

mit dieser Aufgabe getestet werden soll, nämlich das Sprachverständnis, d.h. die Übertragungsfähigkeit einer deutschen Sprachwendung in die englische Sprachwendung. In Abbildung 31 ist der Kreis D voll in das mit durchgehendem Strich gezeichnete Oval hineingezeichnet. Das Oval verdeutlicht das abgeprüfte Konzept.

Die Distraktoren bei dieser Aufgabe sind die Alternativen A, B und C. Alternative A ist ein wenig effektiver Distraktor. Die vorgegebene Übersetzung trifft nur gering den deutschen Satz. Die Schnittmenge mit dem Testkonzept ist sehr klein. B und C sind gute Distraktoren. Am besten wirkt Alternative B; sie kommt zur Hälfte mit dem Testkonzept zur Deckung. Beide Alternativen können den Schüler, der ein nicht ausgereiftes Sprachverständnis im Englischen hat, dazu verleiten, sie anzukreuzen.

Das gestrichelt gezeichnete Oval in Abbildung 31 verdeutlicht das Konzept eines Schülers, der Alternative B als richtig ansieht.

Mehrfachwahlaufgaben sollen extrinsisch mehrdeutig sein, d.h., eine Aufgabe soll für denjenigen mehrdeutig sein, der ein falsches Verständnis des Inhaltes hat, das getestet wird.

Die Aufgaben dürfen aber nicht so konstruiert sein, daß sie intrinsisch mehrdeutig sind. Aufgaben sind intrinsisch mehrdeutig, wenn sie auch für denjenigen mehrdeutig sind, der sich im Testgebiet gut auskennt: m.a.W., wenn die Schnittmenge der richtigen Alternative mit dem Testkonzept nur geringfügig größer ist als die der falschen. Sie werden solche Mehrfachwahlaufgaben kennen; die Alternativen unterscheiden sich nur in Feinheiten der Formulierung. Der Testkonstrukteur hat eine zu feine Unterscheidung zwischen den Alternativen gezogen, so daß das Auffinden der richtigen Antwort zu einem Puzzle wird.

Geringe Trennschärfeindizes bei Mehrfachwahlaufgaben können auf intrinsisch mehrdeutige Aufgaben verweisen. Wenn die im Gesamtttest leistungsstarken Schüler bei einer Aufgabe in der Regel nicht die Alternative wählen, die der Testkonstrukteur als richtig ansieht, dann muß angenommen werden, daß diese Aufgabe Konstruktionsfehler aufweist.

Bei Aufgabe 3, Tabelle 29, ist dies der Fall. Eine Auszählung der Antwortalternativen dieser Aufgabe für die leistungsstarken und leistungsschwachen Schüler mag ein Bild ergeben, wie es Tabelle 31 zeigt. Alternative B bedarf einer näheren Betrachtung. Vermutlich ist die Unterscheidung zwischen Alternative B und C zu geringfügig. Der vorangegangene Unterricht mag nicht ausdrücklich diesen Unterschied hervorgehoben haben. Unter Umständen wird sich herausstellen, daß beide Alternativen, B und C, sinnvolle Lösungen zur Aufgabe sind. In diesem Falle muß noch nachträglich Alternative B mit einem Rohwert bewertet werden. Distraktor A ist unbrauchbar und sollte durch einen anderen Distraktor ersetzt werden. Dagegen arbeitet Distraktor D ausgezeichnet.

Aufgabe 4 wies eine negative Trennschärfe auf (Tabelle 29). Allein eine Analyse des Inhalts der ganzen Testaufgabe und eine Befragung der getesteten Schüler kann darüber Auskunft geben, wie eine Antwortverteilung, wie sie in Tabelle 32 aufgezeichnet ist, zustandekommen konnte.

Aufgabe	Gruppe	Alternative				Ausgelassen	Σ
		A	B	C*	D		
3	Hohe	0	3	4	0	1	8
	Niedrige	0	0	4	4	0	8

* richtige Antwort

Tab. 31:

Antwortverteilung zu Aufgabe 3 für Schüler mit im Gesamttest hohen und niedrigen Rohwerten, vgl. Tabelle 29 (nach J.C. Stanley, K.D. Hopkins, Educational and Psychological Measurement and Evaluation, Englewood Cliffs, New Jersey 1972, S.275).

Aufgabe	Gruppe	Alternative				Ausgelassen	Σ
		A*	B	C	D		
4	Hohe	3	0	1	4	0	8
	Niedrige	7	1	0	0	0	8

* richtige Antwort

Tab. 32:

Antwortverteilung zu Aufgabe 4 für Schüler mit im Gesamttest hohen und niedrigen Rohwerten, vgl. Tabelle 29 (nach J.C. Stanley, K.D. Hopkins, Educational and Psychological Measurement and Evaluation, Englewood Cliffs, New Jersey 1972, S.275).

11.5 Aufgabenanalyse für den Beispieltest Vitamine und Mineralstoffe

Die Zahl der Schüler, die sich dem informellen Leistungstest über Vitamine und Mineralstoffe unterzogen, betrug $N = 30$. Der Gruppenumfang n der im Gesamttest „hohen“ und „niedrigen“ Schüler ergibt sich über $n = N \cdot 0,27$ zu $30 \cdot 0,27 = 8,1 \approx 8$. Somit zählen die acht besten Schüler zur hohen Gruppe (H), die acht schlechtesten zur niedrigen Gruppe (N). Tabelle 33 gibt die Ergebnisse der Aufgabenanalyse wieder.

Aufgabe 1 lösten 5 Schüler der hohen Gruppe; aus Tabelle 2 (S. 16), der Grundauszählung des Testes, läßt sich dies nachprüfen. Zur hohen Gruppe zählen die 8 Schüler von Christa H. bis einschließlich Christa W. (Tabelle 2). Von der niedrigen Gruppe lösten zwei Schüler diese Auf-

gabe. Die niedrige Gruppe reicht in Tabelle 2 von Marlies G. bis einschließlich Thomas B. Die Trennschärfe dieser Aufgabe beträgt dann $D = 0,38$, die Schwierigkeit $p = 0,44$.

Aufgabennummer		Anzahl d. richtigen Antworten	Prozentsatz d. richt. Antworten	Trennschärfe $D = p_H - p_N$	Schwierigkeitsgrad $p = \frac{p_H + p_N}{2}$
1	H	5	$\frac{5}{8} = 0,63$	$0,63 - 0,25 = 0,38$	$\frac{0,63 + 0,25}{2} = 0,44$
	N	2	$\frac{2}{8} = 0,25$		
2	H	6	$\frac{6}{8} = 0,75$	0,62	0,44
	N	1	0,13		
3	H	8	1	0,25	0,88
	N	6	0,75		
4	H	8	1	0,12	0,94
	N	7	0,88		
5	H	6	0,75	0,12	0,69
	N	5	0,63		
6	H	6	0,75	0,12	0,69
	N	5	0,63		
7	H	6	0,75	0,37	0,57
	N	3	0,38		
8	H	7	0,88	0,38	0,69
	N	4	0,50		
9	H	5	0,63	0,63	0,32
	N	0	0		
10	H	7	0,88	0,50	0,63
	N	3	0,38		

Aufgabennummer		Anzahl d. richtigen Antworten	Prozentsatz d. richt. Antworten	Trennschärfe $D = P_H - P_N$	Schwierigkeitsgrad $P = \frac{P_H + P_N}{2}$
11	H	7	0,88	0,50	0,63
	N	3	0,38		
12	H	8	1	0,25	0,88
	N	6	0,75		
13	H	8	1	0,75	0,63
	N	2	0,25		
14	H	8	1	0,87	0,57
	N	1	0,13		
15	H	5	0,63	0,63	0,32
	N	0	0		
16	H	2	0,25	0,12	0,19
	N	1	0,13		
17	H	6	0,75	0,75	0,38
	N	0	0		
18	H	8	1	0,12	0,94
	N	7	0,88		
19	H	8	1	0,25	0,88
	N	6	0,75		
20	H	1	0,13	-0,25	0,26
	N	3	0,38		

Tab. 33:

Aufgabenanalyse für den Beispielttest Vitamine und Mineralstoffe, vgl. Tabelle 2.

Ein Blick auf die Trennschärfeindizes D der Aufgaben in Tabelle 33 zeigt, daß die unten genannten Aufgaben wegen zu geringer Trennschärfe revisionsbedürftig sind. Dies sind Aufgaben mit einem D kleiner als 0,30. Aufgaben mit einem D kleiner als 0,20 sind unterstrichen.

Aufgabe 3	$D = 0,25$
Aufgabe 4	$D = 0,12$
Aufgabe 5	$D = 0,12$
Aufgabe 6	$D = 0,12$
Aufgabe 12	$D = 0,25$
Aufgabe 16	$D = 0,12$
Aufgabe 18	$D = 0,12$
Aufgabe 19	$D = 0,25$
Aufgabe 20	$D = -0,25$

Tab. 34:

Aufgaben des Testes Vitamine und Mineralstoffe mit zu geringer Trennschärfe.

Abbildung 32 vermittelt grafisch einen Eindruck über die Trennschärfe und Schwierigkeit der Testaufgaben.

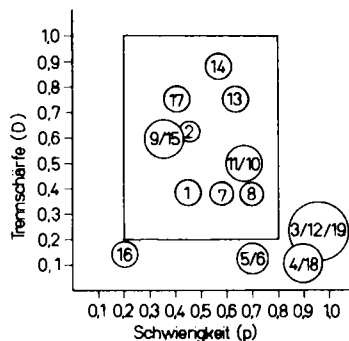


Abb. 32:

Hilfsdiagramm zur Selektion von Testaufgaben im Test Vitamine und Mineralstoffe, ohne Aufgabe 20.

Insgesamt erweisen sich also 9 Aufgaben als revisionsbedürftig. Gespräche mit den Schülern über diese Testaufgaben führten zu einer revidierten Testfassung, wie sie auf den folgenden Seiten wiedergegeben ist³⁹. Die Aufgaben 12, 19 und 20 sind beibehalten worden, obwohl sie von der Trennschärfe her sich als revisionsbedürftig erwiesen haben. Das Gespräch mit den Schülern zeigte, daß der vorangegangene Unterricht nicht deutlich auf die Inhalte dieser drei Testaufgaben abgestellt war; dies galt insbesondere für Aufgabe 20.

Im folgenden Kapitel wird auf die sinnvolle Gestaltung von Testaufgaben näher eingegangen. Spätestens bei der Testrevision ist der Testentwickler wieder auf grundsätzliche Erörterungen über Testaufgaben und ihre Konstruktion angewiesen.

Name:

Klasse:

Datum:

Test: Vitamine und Mineralstoffe

Arbeitsanweisung: Bei vorgegebenen Antworten die zutreffenden (gegebenfalls mehrere) ankreuzen! Freie Stellen mit kurzen Antworten füllen und wo gefordert, eine Begründung angeben!

1. Zu welchen Stoffen zählen wir die Vitamine? Zu den
a) Baustoffen ☐
b) Wirkstoffen ☐
c) Brennstoffen ☐
2. Zu welchen Stoffen zählen wir die Mineralstoffe? Zu den
a) Brennstoffen ☐
b) Wirkstoffen ☐
c) Baustoffen ☐
3. Gib für jedes der beiden folgenden Vitamine an, zu welcher Gruppe sie gehören!
Vitamin A _____
Vitamin K _____
4. Warum unterscheiden wir zwei Gruppen von Mineralstoffen? _____

5. Welches Vitamin ist vor allem im Brot und in Getreideerzeugnissen zu finden? _____
6. Wo finden wir vorwiegend fettlösliche Vitamine? _____

7. Gib ein Beispiel für eine Vitaminmangelkrankung an! _____

8. Wodurch wird die Wirksamkeit der Vitamine beeinträchtigt?
a) Wasser ☐ c) Luft ☐ e) Trockenheit ☐
b) Kälte ☐ d) Hitze ☐ f) Licht ☐
9. Wie lange kann man ein Nahrungsmittel mit einem hohen Anteil fettlöslicher Vitamine wässern, ohne daß der Vitamingehalt beeinträchtigt wird? Begründe Deine Antwort kurz!
a) nur kurz _____ ☐
b) längere Zeit _____ ☐
c) unbegrenzt _____ ☐
10. Welchen chemischen Stoff bezeichnet man mit Vitamin C? _____

11. In welcher Größenordnung liegt der tägliche Bedarf an Vitaminen? (keine Zahlenangabe) _____

12. Warum gehören die Vitamine und Mineralstoffe für den Körper zu den essentiellen Stoffen? _____

13. Gib 3 Dir bekannte Mengenelemente an!
1. _____ 2. _____ 3. _____
14. Gib 3 Dir bekannte Spurenelemente an!
1. _____ 2. _____ 3. _____
15. Welches Nahrungsmittel ist besonders reich an Mineralstoffen? (vor allem an Ca) _____
16. Welche Funktion haben die Mengenelemente? _____

17. Welche Funktion haben die Spurenelemente? _____

18. In welchen Nahrungsmitteln finden wir Mineralstoffe gelöst?
Gib 3 Beispiele an!
1. _____
2. _____
3. _____
19. Warum sollte man ungeschälten Reis geschältem Reis vorziehen?

20. Wie nennt man die Vorstufe von Vitaminen? _____

12. Testaufgaben und ihre Konstruktion

Testaufgaben werden unterschieden nach der Art und Weise, wie die Antwort verlangt wird. Es gibt Aufgaben mit gebundener und Aufgaben mit freier Beantwortung. Bei der gebundenen Beantwortung ist vom Testentwickler der Rahmen festgelegt, in dem der Testbearbeiter seine Antwort geben kann. Bei der freien Beantwortung läßt sich vom Testbearbeiter selbst die Antwort ohne jegliche Einschränkung formulieren.

Zu den Aufgaben mit gebundener Beantwortung zählen Richtig-Falsch-Aufgaben, Mehrfachwahlaufgaben, Zuordnungsaufgaben und Umordnungsaufgaben. Unter Aufgaben mit freier Antwortformulierung fallen die Ergänzungsaufgaben und Kurzantwortaufgaben (Kurzaufsatzaufgaben).

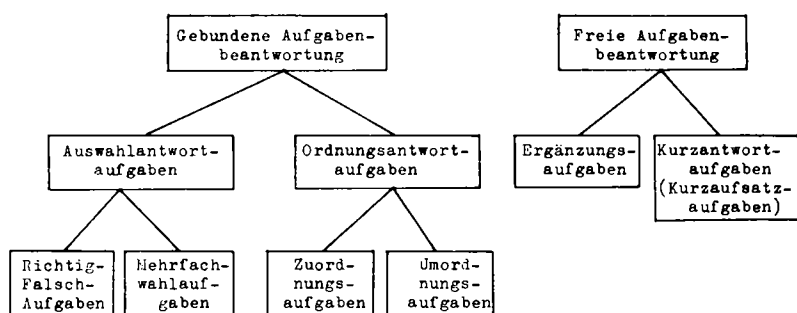


Abb. 33:
Übersicht über die wichtigsten Aufgabentypen

12.1 Richtig-Falsch-Aufgaben

i) *Aufbau:* Diese Aufgaben bestehen aus Feststellungen, die der Testbearbeiter als richtig oder falsch zu kennzeichnen hat. Andere Formen verlangen, „Ja“ oder „Nein“ zu kennzeichnen oder fordern auf, zuzustimmen oder abzulehnen („stimmt“, „stimmt nicht“). Beispiele:

Sind die folgenden Aussagen richtig, kreuze "R" an, sind sie falsch, kreuze "F" an!

	R	F
- Die Standardabweichung ist ein Maß für die Variabilität von Testrohwertsätzen	☐	☐
- Die Erde kreist um die Sonne	☐	☐
- $\frac{4}{5}$ ist mehr als $\frac{5}{6}$	☐	☐

ii) *Konstruktionshinweise:*

1. Die Feststellungen müssen eindeutig richtig oder falsch sein. Es muß ausgeschlossen sein, daß mit einer Interpretation die Aussage richtig sein mag, und mit einer anderen gleich plausiblen Interpretation die Aussage auch falsch sein mag, z. B.: „Strom fließt im Verbraucher von minus nach plus“ ist richtig, wenn von der physikalischen Stromrichtung ausgegangen wird, ist aber falsch, wenn konventionsgemäß die technische Stromrichtung zugrunde gelegt wird. Erst der Zusatz („Technischer) Strom fließt im Verbraucher von minus nach plus“ macht diese Aussage eindeutig falsch.
2. Ungewollte Hinweise auf die richtige Antwort sind zu vermeiden. Zusätze im Aussagesatz wie „immer“, „alle“, „niemals“, „keine“, „nichts“ deuten auf falsche Feststellungen hin. Worte wie „kann“, „mag“, „manchmal“, „einige“, „häufig“, „oft“, „im allgemeinen“ und „in der Regel“ sind Anzeiger für richtige Feststellungen. Z. B. deutet die Formulierung „Eine Relation ist niemals eine Funktion“ schon darauf hin, daß diese Feststellung falsch ist. Umgekehrt wird die Formulierung „Eine Relation kann eine Funktion sein“ eher dazu hinführen, diese Aussage als richtig zu beurteilen.
3. Die Anzahl der falschen und richtigen Feststellungen sollten in etwa gleich groß sein. Die falschen und richtigen Behauptungen sollten dem Zufallsprinzip nach aufeinander folgen.
4. Negative Aussagen und doppelte Verneinungen sollten vermieden werden: z. B. nicht „Der Median ist kein Maß der zentralen Tendenz eines Rohwertsatzes“ sondern „der Median ist ein Maß der zentralen Tendenz eines Rohwertsatzes“ (R/F).
5. Die Aussagesätze sollten kurz, prägnant und verständlich sein.
6. Die Aussagesätze sollten in Gruppen angeordnet sein. Der Testbearbeiter wendet eine geringere Wahrnehmungskraft auf, wenn die Sätze in Gruppen zu fünf oder sechst mit doppeltem Abstand zwischen den Gruppen angeordnet sind.

iii) *Vorteile:* Richtig-Falsch-Aufgaben sind leicht zu konstruieren und vom Testbearbeiter schnell zu lösen. Die richtige Antwort steht objektiv fest. Die Auswertung ist einfach. Ein Test mit Richtig-Falsch-Aufgaben kann für die gleiche Bearbeitungszeit wesentlich mehr Testaufgaben beinhalten als ein Test mit anderen Aufgabenformen. Das heißt, ein Richtig-Falsch-Test kann ein weites Wissensgebiet in kurzer Bearbeitungszeit abtesten.

iv) *Nachteile:* Aufgaben mit den Alternativen „Richtig“, „Falsch“ prüfen in der Regel nur Faktenwissen ab. Es ist außerordentlich schwer, diese Aufgaben so zu konstruieren, daß sie auch Lernziele abtesten, die

höheren kognitiven Stufen wie Verstehen und Anwendung zuzuschreiben sind.

Der gewichtigste Einwand gegen die Richtig-Falsch-Aufgaben ist die hohe Ratewahrscheinlichkeit für richtige Lösungen. Der Testbearbeiter kann bei willkürlichem Ankreuzen im Durchschnitt damit rechnen, daß er zu 50% richtige Antworten erhält.

Die Nachteile der Richtig-Falsch-Aufgaben führen zu der Überlegung, solche Aufgaben in einem Test nur wenig einzusetzen. Sie bieten sich dann an, wenn andere Testaufgabenarten nicht geeignet erscheinen. So kann es manchmal schwierig oder sogar unmöglich sein, für eine Mehrfachwahlaufgabe mehrere plausible Antworten zu finden. In diesem Fall wird man auf eine einfache Richtig-Falsch-Aufgabe zurückgreifen.

12.2 Mehrfachwahlaufgaben⁴⁰

i) *Beschreibung:* Die Mehrfachwahlaufgabe bietet in der Regel vier oder fünf Antwortmöglichkeiten an, von denen nur eine die richtige oder die ausgesprochen beste Lösung darstellt. Dieser Aufgabentyp besteht formal aus einem sog. „Stamm“, in dem ein Problem in Form einer Frage oder Feststellung vorgegeben wird. An diese Grundinformation schließen sich die vorgegebenen Antworten als „Alternativen“ an. Eine der Antworten ist in der Regel richtig; die anderen Antworten, die „Distraktoren“ sollen mehr oder weniger plausible Alternativen zur richtigen Antwort darstellen. Sie sollen einen Teil der Signierungen von der richtigen Lösung auf sich hinziehen (distrahieren).

Beispiel 1:

Das Elektron kreist mit hoher Geschwindigkeit um den Atomkern. Von welcher Art ist die Kraft, die der Fliehkraft entgegenwirkt und das Elektron auf seiner Bahn hält? Es ist eine

- a) elektromagnetische Anziehungskraft.
- b) Schwerkraft (Gravitationskraft).
- c) Kohäsionskraft.
- d) elektrostatische Anziehungskraft⁴¹.

Das o.g. Beispiel gehört zum „Richtige-Antwort“ Typ einer Mehrfachwahlaufgabe. Nur die Alternative d ist richtig, alle anderen Alternativen sind falsch. Bei den Mehrfachwahlaufgaben ist auch denkbar, daß aus vorgegebenen Antworten die beste ausgewählt wird. Vom „Beste-Antwort“ Typ ist die folgende Mehrfachwahlaufgabe, bei der der Testbear-

beiter jenen englischen Satz auswählen soll, der die beste Übersetzung des deutschen Satzes darstellt⁴².

Beispiel 2:

Das Grundstück gehört immer noch Vater.

- a) Father always hears good pieces.
- b) The piece of ground still belongs to Father.
- c) The property still hears to Father.
- d) The property still belongs to Father.

Die Konstruktion von Mehrfachwahlaufgaben des Typs „Beste-Antwort“ erfordert größte Sorgfalt sowie viel Aufwand und stellt an den Testentwickler hohe Anforderungen.

ii) *Konstruktionshinweise:*

1. Der Aufgabenstamm sollte *ein* zentrales Problem enthalten, alle notwendigen Informationen geben und frei von irrelevantem Material sein. Wörter, die in jeder Auswahlantwort wiederholt werden müssen, gehören in den Stamm der Aufgabe (vgl. zu diesen Anforderungen das Beispiel 1).
2. Negativ formulierte Aufgabenstämme sollten vermieden werden. Falls jedoch die Negativform benutzt werden muß, ist dies durch Hervorhebung (Unterstreichung, gesperrtes Schreiben) deutlich zu machen. Z. B.: „Welchen spezifischen Widerstand darf Kupfer für elektrotechnische Zwecke nach den VDE-Bestimmungen *nicht* überschreiten?“ – Sofern es möglich ist, sollten Mehrfachwahlaufgaben mit negativ formulierten Aufgabenstämmen zu einem Block zusammengefaßt werden.
3. Die Distraktoren müssen plausibel sein, vgl. hierzu Kap. 11.4 Distraktorenanalyse.
4. Die Auswahlantworten müssen alle grammatikalisch mit dem Stamm der Aufgabe übereinstimmen. Der Testbearbeiter darf nicht allein aufgrund der grammatikalischen Richtigkeit einer Alternative einen Hinweis auf die Lösung der Aufgabe bekommen.

Beispiel 3:

Welche der folgenden Männer waren berühmte Physiker?

- | | |
|----------------------|---|
| - van Gogh | A |
| - Liebermann | B |
| - Einstein und Hertz | C |
| - Ebert | D |

Die Pluralformen „welche“ und „waren“ weisen schon auf die richtige Antwort C hin. Die Distraktoren A, B und D müssen noch um jeweils einen Namen ergänzt werden.

5. Die Position der richtigen Antwort sollte dem Zufallsprinzip überlassen werden: mit anderen Worten, die richtige Antwort darf nicht immer am selben Platz stehen.
6. Jede Aufgabe sollte zumindest über vier Auswahlantworten verfügen.
7. Die Aufgaben dürfen sich nicht überschneiden, d.h. der Aufgabestamm einer Aufgabe darf nicht die Antwort einer folgenden oder vorhergehenden Aufgabe enthalten bzw. erleichtern.
8. Wörter wie „immer“ und „niemals“, die ungewollt einen Hinweis auf die richtige Antwort geben, sind zu vermeiden (vgl. Hinweis 2 bei den Richtig-Falsch-Aufgaben).
9. Die Länge, Ausführlichkeit oder Exaktheit in der Formulierung einer Auswahlantwort darf nicht in einem Zusammenhang mit der richtigen Antwort stehen. Trifft dies zu, wird dem Testbearbeiter unbeabsichtigt ein Hinweis auf die zutreffende Lösung gegeben.

Beispiel 4:

Ein Atom ist

- a) eine Mischung.
- b) eine Verbindung.
- c) ein Molekül.
- d) der kleinste Baustein der Materie, bestehend aus Nukleonen, die von Elektronen umkreist werden.

Der Testverfahren wird auch ohne jede Kenntnis über den Aufbau eines Atoms kaum Zweifel daran haben, daß Alternative d) richtig ist. Diese Antwort sticht gegenüber den anderen hervor.

iii) *Vorteile:* Die Mehrfachwahlaufgaben sind, wie die Richtig-Falsch-Aufgaben, ökonomisch. In relativ kurzer Zeit lassen sich viele Aufgaben bearbeiten und auswerten. Ein weites Stoffgebiet läßt sich abdecken.

Die Ratewahrscheinlichkeit für die zutreffende Antwort nimmt mit zunehmender Zahl der Auswahlantworten ab. Bei vier Auswahlantworten beträgt sie lediglich 25 %.

Mehrfachwahlaufgaben eignen sich nicht nur für das Abprüfen von Faktenwissen. Sie eignen sich auch besonders für die Überprüfung von Lernzielen, die höhere geistige Anforderungen stellen, wie z.B. Verstehen, Anwenden oder Analysieren. Die im folgenden aufgeführte Mehrfachwahlaufgabe, gestellt nach der unterrichtlichen Erarbeitung der Prinzipien einer Dampferzeugung, möge dies verdeutlichen.

Beispiel 5:

Wenn die Eruption eines Geysirs beginnt, fließt zunächst heißes Wasser über den Krater, und danach erfolgt ein Dampfausbruch, gemischt mit heißem Wasser. Das erste Ausfließen von heißem Wasser unterstützt die Dampfproduktion, weil

- a) weniger Wasser erhitzt werden muß.
- b) mehr Wasser in den Spalt aus dem umgebenden Stein einsickern kann.
- c) je höher der Druck, desto größer die Dampferzeugung.
- d) je niedriger der Druck, desto niedriger die Temperatur, bei der Dampf erzeugt wird.
- e) die Temperatur des überfließenden Wassers notwendigerweise unter 100 Grad liegt⁴³.

Wenn angenommen werden kann, daß dem Testbearbeiter die Prinzipien der Dampfproduktion bekannt sind, muß er hier das Erlernte auf eine realistische Situation mit einem neuen Gesichtspunkt anwenden⁴⁴.

iv) Nachteile: Die Konstruktion sinnvoller Distraktoren bei den Mehrfachwahlaufgaben ist oftmals schwierig. Man kommt leicht zu irrelevanten Auswahlantworten, die vom Testbearbeiter sofort ausgeschaltet werden können. Eine Mehrfachwahlaufgabe reduziert sich dann schnell zu einer „Einfachwahlaufgabe“. – Die Konstruktion der Mehrfachwahlaufgabe ist eher von einem Team als von einer Person allein zu bewältigen.

Zur Überprüfung kreativer Leistungen sind die Mehrfachwahlaufgaben bedingt geeignet. Die Aufeinanderfolge einzelner Denkschritte bei der Lösung eines Problems läßt sich nicht festhalten. Wird ein mathematisches Problem in Form der Mehrfachwahlaufgabe gestellt, so kann nur das richtige oder falsche Ergebnis ermittelt werden.

12.3 Zuordnungsaufgaben

i) Beschreibung: Die Zuordnungsaufgabe besteht aus zwei Spalten. Eine Zeile der ersten Spalte (Aufgabe) muß einer Zeile der zweiten Spalte (Antwortalternative) zugeordnet werden. In der einfachen Form geht die Zuordnung auf, d. h. die Zahl der Antwortalternativen ist gleich groß wie die Zahl der Aufgaben. In der Regel enthält jedoch die Zuordnungsaufgabe mehr Antwortalternativen als Aufgaben, um die Ratewahrscheinlichkeit bei der Lösung herabzusetzen.

Beispiel 1:

Ordne die Hauptstädte 1 bis 4 den Ländern a - h richtig zu.
(Die Zuordnung geht nicht auf!)

	___a	Frankreich	
	___b	Griechenland	
1	Moskau	___c	Irland
2	Rom	___d	Italien
3	Madrid	___e	Norwegen
4	Oslo	___f	Schweden
	___g	Sowjetunion	
	___h	Spanien	

ii) Konstruktionshinweise:

1. Die Aufgaben sollen aus einem eng umgrenzten, deutlich abgesteckten Wissensgebiet und nicht mehreren Wissensgebieten entstammen. Ist dies nicht der Fall, lassen sich die Antwortalternativen vom Testbearbeiter nach den abgefragten Wissensblöcken sortieren. Innerhalb der Wissensblöcke kann dann die Zahl der Alternativen bis auf eine geschrumpft sein, so daß der Testbearbeiter über seine Test Erfahrung anstatt aufgrund seines Wissens zur richtigen Antwort kommt.
2. Die Listen der Aufgaben und Antwortalternativen sollten unterschiedlich lang sein: die letztere möglichst um 50 % länger als die erste. Die Ratewahrscheinlichkeit läßt sich so reduzieren.
3. Es muß durch eine sorgfältige Testinstruktion sichergestellt werden, daß der Testbearbeiter genau versteht, wie zugeordnet werden soll. Hinweise, daß die Zuordnung nicht aufgeht, oder daß u. U. eine Alternative doppelt verwendet wird, sollten gegeben werden.
4. Die Antwortalternativen sollten in systematischer Reihenfolge angeordnet sein, z. B. nach dem Alphabet oder nach einer zeitlichen Folge. Dem Testbearbeiter erleichtert dies die Übersicht.
5. Die Aufgaben sollen links angeordnet sein, versehen mit arabischen Ziffern, die Antwortalternativen rechts, versehen mit Buchstaben (oder umgekehrt, die Aufgaben mit Buchstaben und die Alternativen mit Ziffern).
6. Die Aufgaben und Alternativen einer Zuordnungsaufgabe sind auf einer Blattseite anzuordnen. Das Vor- und Zurückblättern einer Seite, um die richtige Zuordnung zwischen Aufgabe und Antwortalternative zu finden, ist verwirrend und zeitraubend.

iii) Vorteile: Mit Zuordnungsaufgaben kann auf kleinem Raum eine Fülle von Information abgeprüft werden. Diese Aufgabenform eignet sich besonders für das Überprüfen von Faktenwissen, wenn Zuordnungen assoziativer Art geprüft werden sollen. Beispiele dafür sind: Ereignisse und

Daten, Ereignisse und Personen, Ereignisse und Orte, Begriffe und Definitionen, Personen und ihre Werke, Autoren und Buchtitel, Fremdwörter und deutschsprachige Synonyme, Regeln und Beispiele, Werkzeuge und ihre Verwendung.

iv) *Nachteile:* Zuordnungsaufgaben sind weitgehend auf die Überprüfung von Faktenwissen eingegrenzt. Es dürfte außerordentlich schwierig sein, Zuordnungsaufgaben zu konstruieren, die Verstehensleistungen, Transferleistungen oder sogar analytische Denkvorgänge abzuprüfen imstande sind. – Problematisch kann die Konstruktion einer ausreichenden Zahl von plausiblen Antwortalternativen sein, zumal die Zahl der Antwortalternativen doppelt so groß sein soll wie die der Aufgaben.

12.4 Umordnungsaufgaben

i) *Beschreibung:* Bei Umordnungsaufgaben sind dem Testbearbeiter eine Reihe von Begriffen, Wörtern, Zahlen, Symbolen o.ä. vorgegeben, die nach einer bestimmten Gesetzmäßigkeit in eine sinnvolle Reihenfolge umzuordnen sind: z. B. sind Flüsse nach ihrer Länge oder Wörter nach einer sinnvollen Satzstellung zu ordnen.

Beispiel 1:

Tochter sagt zu ihrer Mutter:

You	been	have	should	listening	me	to
A	B	C	D	E	F	G

Richtige Ordnung:
(Trage nur die Buchstaben ein)

1. ____ 2. ____ 3. ____ 4. ____ 5. ____ 6. ____ 7. ____

Beispiel 2:

Ordne die gegebenen Länder nach ihrer jährlichen Fördermenge für Kohle.

(1. = größte Fördermenge, 5. = kleinste Fördermenge)

(Trage nur die Buchstaben ein)

A) Kanada	1. ____
B) Sowjetunion	2. ____
C) USA	3. ____
D) Japan	4. ____
E) Bundesrepublik Deutschland	5. ____

ii) *Konstruktionshinweise:*

1. Die Liste der Begriffe, Wörter, Zahlen, Symbole o.ä. sollte nicht zu lang sein (höchstens 8); der Testbearbeiter verliert sonst leicht die Übersicht.
2. Es muß sichergestellt werden, daß eine eindeutige Ordnung möglich ist.
3. Die zu ordnenden Begriffe, Wörter, Zahlen, Symbole o.ä. sind mit Buchstaben und die Freiplatzvorgaben für die sinnvolle Ordnung mit arabischen Ziffern zu versehen (oder umgekehrt). Der Aufgabenteil muß vom Lösungsteil der Umordnungsaufgabe deutlich getrennt sein, aber räumlich auf etwa gleicher Höhe angeordnet werden (vgl. Beispiel 1 u. 2).

iii) *Vorteile:* Umordnungsaufgaben sind einfach zu konstruieren. Sie bieten sich dort an, wo Wissen, das eine spezielle Ordnung aufweist, ökonomisch überprüft werden soll. Bei geschickter Konstruktion dürften sich mit diesem Aufgabentyp auch Denkvollzüge abprüfen lassen, die über den Reproduktionsbereich hinausgehen.

iv) *Nachteile:* Umordnungsaufgaben lassen sich nur zum Überprüfen der wenigen Sachverhalte einsetzen, die sinnvollerweise eindeutig in einer Ordnung bzw. nach einer Reihenfolge darstellbar sind.

Problematisch ist die Auswertung. Der Testkonstrukteur hat eindeutig festzulegen, ab wann eine Umordnungsaufgabe als gelöst gilt: ob alle Glieder in der Reihe richtig stehen müssen oder ob es ausreicht, wenn 80% der Glieder richtig angeordnet sind, ob Teillösungen mit Punkten versehen werden und wenn ja, mit wie vielen. Schon bei $N = 4$ zu ordnenden Vorgaben gibt es 24 verschiedene Anordnungsmöglichkeiten ($4 \cdot 3 \cdot 2 \cdot 1 = 24$; allgemein: $N \cdot (N-1) \cdot (N-2) \dots (1)$ mögliche Anordnungen für N Vorgaben). Es ist offensichtlich, welche Schwierigkeiten hier bei der Auswertung auftreten können.

12.5 Ergänzungsaufgaben

i) *Beschreibung:* Ergänzungsaufgaben bestehen aus einem Satz, in dem bestimmte wichtige Wörter, Zahlen, Zeichen, Symbole o.ä. ausgelassen sind. An ihrer Stelle befindet sich eine Lücke, die der Testbearbeiter auszufüllen hat. Geschieht dies, ist der Satz vollständig „ergänzt“. Der Satz kann eine oder zwei Lücken enthalten. In der Regel wird es bei einer Lücke belassen.

Beispiele:

1. Kolumbus entdeckte im Jahre 1 Amerika. 1 _____
2. Ein Mann läuft nach Westen, wendet sich dann
um 90 Grad nach rechts. Er geht nun in 2
Himmelsrichtung weiter. 2 _____
3. Das "Kap der guten Hoffnung" wurde zuerst von
3 umsegelt. 3 _____

ii) Konstruktionshinweise:

- 1) Die Ergänzungsaufgabe ist so zu formulieren, daß nur eine bestimmte Antwort möglich ist. Dies wird erreicht, indem zuerst eine direkte, eindeutig gehaltene Frage aufgestellt wird, die daraufhin in einen vollständigen Antwortsatz umformuliert wird. Dabei werden alle wesentlichen Worte der Frage benutzt und an der Stelle der abzufragenden Information wird eine Lücke gelassen.

Beispiel:

- Die vage, uneindeutige Frage: „Wo schlug Martin Luther seine Thesen an?“ führt zu der *schlechten* Ergänzungsaufgabe:

1. Martin Luther schlug seine Thesen 1 an. 1 _____

- Die eindeutige Frage: „In welcher Stadt schlug Martin Luther seine Thesen an die Kirchentür an?“ führt zu der Ergänzungsaufgabe:

2. Martin Luther schlug seine Thesen in der Stadt
1 an die Kirchentür an. 1 _____

Bei der ersten Fassung ergeben die Angabe der Stadt, des Gebietes, des Landes – oder einfach „Kirchentür“ – legitime Antworten. Die zweite Fassung ist eindeutig; hier gibt es nur eine Antwort.

- 2) Übermäßig verstümmelte Aussagesätze sollten vermieden werden. Sind zu viele Schlüsselwörter ausgelassen, ist nicht mehr deutlich, welche Wörter eigentlich gesucht sind.

Beispiel:

1. Der 1 einer Person ist der 2 aus der 3 1 _____
vom Mittelwert und der 4 . 2 _____
3 _____
4 _____

In dieser Form wird die Ergänzungsaufgabe zum Rätselspiel. Eindeutiger ist die Formulierung:

2. Der Standardwert z einer Person ist der Quotient aus der Rohwertabweichung vom Mittelwert und der 1 . 1 _____

- 3) Grammatikalische Hinweise auf die richtige Antwort sind zu vermeiden.

Beispiel:

1. Den Atomkern bilden 1 . 1 _____

Die Formulierung ohne einen grammatikalischen Hinweis:

2. Der Atomkern besteht aus 1 . 1 _____

In der ersten Fassung ist durch die Pluralform „bilden“ angezeigt, daß der Atomkern aus mehr als einem Baustein besteht. In der zweiten Fassung ist dies offen gelassen.

- 4) Die Lücken für die Antworten sollten innerhalb einer Ergänzungsaufgabe und von Aufgabe zu Aufgabe immer gleich lang sein. Variiert die Länge der Lücken mit der Länge der richtigen Antwort, erhält der Testbearbeiter einen Hinweis auf die Lösung.
- 5) Die Aufgaben sind so anzulegen, daß auf der rechten Seite des Testbogens in einer Spalte die Antworten eingetragen werden können (siehe hierzu die oben aufgeführten Beispiele).
- 6) Für die Auswertung ist ein Antwortschlüssel vorzubereiten, der alle akzeptablen Antworten enthält. Es ist zwar wünschenswert, daß für jede Lücke nur eine Antwort richtig ist, dies wird aber nicht immer möglich sein.

iii) *Vorteile:* Ergänzungsaufgaben sind leicht zu konstruieren. Die Ratenwahrscheinlichkeit ist herabgesetzt: es sind weniger Anhaltspunkte zum Raten auffindbar als bei vorgegebenen Antwortalternativen.

iv) *Nachteile:* Mit Ergänzungsaufgaben lassen sich in der Regel nur einfache Sachverhalte (Faktenwissen) abprüfen. Probleme können sich bei der Auswertung einstellen. Ist die Aufgabe nicht exakt formuliert, ergeben sich unerwünschte Antworten, die aber auch als gültig angesehen werden können.

Bei Ergänzungsaufgaben besteht die Gefahr, daß mit diesen Aufgaben mehr die allgemeine Intelligenz oder Lesefähigkeit, als der eigentlich beabsichtigte Testinhalt abgeprüft wird. Dies trifft zu, wenn der Testbearbeiter erst herausfinden muß, was die eigentliche Fragestellung der Aufgabe ist (siehe oben).

12.6 Kurzantwortaufgaben (Kurzaufsatzaufgaben)

i) *Beschreibung:* Auf einem ähnlichen Prinzip wie die Ergänzungsaufgaben beruhen die Kurzantwortaufgaben. Der Unterschied besteht darin, daß hier nicht ein unvollständiger Satz angeboten wird, sondern daß eine Frage oder eine Aufforderung gestellt wird. Die Antwort kann aus nur einem Wort, einem Satz oder auch aus mehreren Sätzen bestehen. Trifft letzteres zu, wird die Kurzantwortaufgabe zur Kurzaufsatzaufgabe. Der Übergang ist fließend.

Beispiele:

1. In welchem Kontinent fließt der Amazonas? 1 _____
2. Gib ein Beispiel für eine Vitaminmangelkrankung an! 2 _____
3. Wie lautet der Satz des Pythagoras (in Worten)? 3 _____

4. Warum gehören die Vitamine und Mineralstoffe für den Körper zu den essentiellen Stoffen? 4 _____

ii) *Konstruktionshinweise:*

- 1) Die Frage oder Aufforderung sollte so formuliert sein, daß die geforderte Antwort kurz ausfällt und daß nur eine einzige Antwort die richtige ist (vgl. Konstruktionshinweis 1 für die Ergänzungsaufgabe).
- 2) Die Freiplätze zum Eintragen der Antworten sollten gleich lang sein und möglichst am rechten Rand der Testseite aufgeführt sein, um die Auswertung zu erleichtern. Wird die Kurzantwortaufgabe jedoch mehr zur Kurzaufsatzaufgabe, läßt sich diese Forderung allerdings nicht mehr sinnvoll erfüllen (vgl. die oben aufgeführten Beispiele).
- 3) Ein Antwortschlüssel ist vorzubereiten, der alle zu akzeptierenden möglichen Antworten enthält. Gerade bei den Kurzaufsatzaufgaben kann es für die Lösung mehr als eine zulässige Interpretation der gestellten Frage geben.

iii) *Vorteile:* Die Kurzantwortaufgaben sind, wie die Ergänzungsaufgaben, verhältnismäßig leicht zu konstruieren. Die Ratewahrscheinlichkeit ist herabgesetzt. Kurzantwortaufgaben, die mehr zu Kurzaufsatzaufgaben werden, ermöglichen die Erfassung von kognitiven Leistungen,

die über den Reproduktionsbereich hinausgehen, wie Verstehen und Anwendung. Der Testbearbeiter muß bei diesen Aufgaben eine eigenständige, komplexe Leistung erbringen.

iv) *Nachteile:* Kurzantwortaufgaben im Sinne von Kurzaufsatzaufgaben erfordern eine lange Bearbeitungszeit. Erhebliche Schwierigkeiten können sich bei der Auswertung ergeben. Besonders bei den Kurzaufsatzaufgaben können sich Antworten einstellen, die vom Testkonstrukteur so nicht vorhergesehen worden sind, andererseits aber auch nicht falsch sind.

Schlußbemerkung

Das Erstellen von Testaufgaben ist eine schöpferische Tätigkeit. Häufigkeitsverteilungen, Mittelwerte, Standardabweichungen, lineare oder nichtlineare Rohwerttransformationen in T- oder Z-Werte, Reliabilitätskoeffizienten und Aufgabenanalysen sind statistisch-rechnerische Verfahren: sie bieten Kontrollmöglichkeiten, vermitteln Hinweise, sind aber keine Garantie für gute Aufgaben. Die statistisch-rechnerischen Verfahren erübrigen nicht, daß Testaufgaben sorgfältig, phantasie reich und in sich logisch entworfen werden müssen. – Dies gelingt am ehesten in einem Team.

13. Notengebung

13.1 Umsetzung der Rohwerte in Noten

Es wird immer einer pädagogischen Entscheidung vorbehalten bleiben, welche Rohwertzahl in einem Test noch mit einem „ausreichend“ zu versehen ist bzw. welche bereits eine „mangelhafte“ Leistung ausweist. Im Beispielttest Vitamine und Mineralstoffe konnten maximal 20 Rohwerte erreicht werden. Erhalten nun die Schüler, die 7 Rohwerte erzielten, die Note 5? Oder zählen zu den Schülern mit mangelhafter Leistung auch jene, die 8, 9 oder sogar 10 Rohwerte erzielten? Die statistisch-rechnerischen Möglichkeiten der Notenfindung, wie sie im folgenden vorgestellt werden, können nur eine Hilfe zur Entscheidungsfindung geben. Sie nehmen die Entscheidung dem einzelnen Lehrer nicht ab. Hierauf sei vorab nachdrücklich hingewiesen.

Vor der Durchführung eines vom Lehrer selbst erstellten Tests, lassen sich die Noten einer bestimmten Anzahl richtig gelöster Aufgaben (Rohwerten) zuordnen. Dieses Verfahren hat allerdings einen schwerwiegenden Nachteil: die Schwierigkeit der einzelnen Aufgaben kann vom Lehrer nicht hinreichend genau geschätzt werden. Es kann sich herausstellen, daß der Test insgesamt zu schwer oder zu leicht war. Der Lehrer kann dann durch seine Beurteilungsvorgabe gezwungen werden, besonders hart bzw. besonders großzügig zu urteilen. Hat sich der Test als zu schwer erwiesen, wird er in der Regel nachträglich korrigierend eingreifen müssen.

Ein anderes Verfahren der Notenfindung besteht darin, sich auf die Rohwerte auszurichten, die in der Klasse gewonnen wurden. Die Notengebung wird hier an dem Mittelwert und an der Verteilung einer Schulklasse ausgerichtet, d. h. also an einem klasseninternen Bezugssystem. Es besteht die Gefahr, daß das klasseninterne Bezugssystem dem einzelnen Schüler und den Leistungsanforderungen des Lehrers u. U. nicht gerecht wird. Diese Gefahr wird reduziert, wenn Schüler mehrerer Schulklassen vergleichbaren Unterrichts die Bezugsgruppe bilden.

Sofern die Bezugsgruppe groß ist, lassen sich sinnvoll die Kenngrößen Mittelwert $\bar{x}$ und Standardabweichung s der Rohwertverteilung der Bezugsgruppe bestimmen. Über $\bar{x}$ und s lassen sich die Schulnoten (SN) als Standardwerte über folgende Formel ausdrücken⁴⁵.

$$SN = 3 - z = 3 - \frac{(x - \bar{x})}{s}$$

Diese Schulnotenskala ist fünfstufig. Der Mittelwert liegt bei der Note 3, die Standardabweichung beträgt eine Notenstufe (Abbildung 34). 68% aller getesteten Personen erhalten eine Leistung im Notenbereich von 2,0

bis 4,0 ($z = -1$ bis $z = +1$). Jeweils 16 % erhalten Noten, die über 2,0 bzw. unter 4,0 liegen. Diese Schulnotenskala (SN) findet Verwendung bei standardisierten Schulleistungstests.

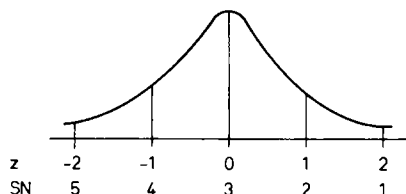


Abb. 34:
Schulnoten (SN) und zugehörige z-Werte der Standardnormalverteilung.

Ein anderes Notenverteilungsschema stellt die hier bereits vorgestellte Stasix-Skala dar (vgl. S. 70). Die Stasix-Skala sieht 6 Abschnitte, d. h. Noten von 1–6 vor. Auf einen mittleren 50 % Bereich links und rechts des Mittelwertes $z = 0$ entfallen die Noten 3 und 4 (Tabelle 35).

Note	Prozent der Fälle	Prozentrang (PR)	z-Wert
1	9	100	$> +1,34$
2	16	91	$+1,34$
3	25	75	$+0,67$
4	25	50	$0,00$
5	16	25	$-0,67$
6	9	< 9	$< -1,34$

die Note 3 liegt zwischen PR 50 und 75, die sind 25% der Fälle

die Note 3 liegt zwischen $z=0,00$ und $z=0,67$

Tab. 35:
Notenverteilung nach der Stasix

Eine weitere Einteilung des „grading on the curve“ sieht für die Note 3 den z-Wertbereich von $z = -0,5$ bis $z = +0,5$ vor. Ausgehend von diesem mittleren Abschnitt folgen nach rechts und links in gleich großen Abschnitten die Noten 1 und 2 bzw. 4 und 5 (Abbildung 35). Die Note 3 entfällt auf 38 %, die Noten 2 und 4 auf je 24 % und die Noten 1 und 5 auf je 7 % der getesteten Schüler⁴⁶.

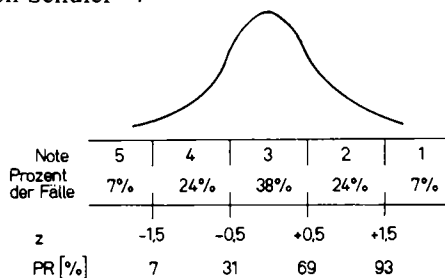


Abb. 35:
Notenverteilung bei einem mittleren z-Abschnitt von $z = -0,5$ bis $z = +0,5$

Von dem Gedanken einer Drittel-Einteilung geht das in Abbildung 36 dargestellte Notenverteilungsschema aus⁴⁷. Auf die Noten 1 und 2, auf die Note 3 und auf die Noten 4 und 5 entfallen jeweils ein Drittel aller getesteten Schüler.

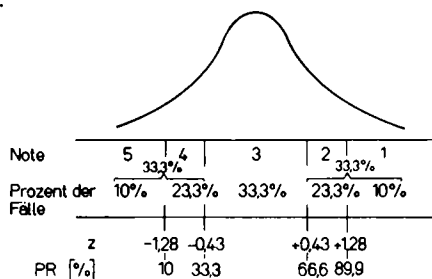


Abb. 36:

Notenverteilung nach einer Drittel-Einteilung: $1/3 = \text{Note 1} + 2$, $1/3 = \text{Note 3}$, $1/3 = \text{Note 4} + 5$.

Bei den hier vorgestellten Notenverteilungsmodellen führen immer zwei Wege über die Rohwerte zu den Noten:

- a) Sie gehen von den z-Werten der Rohwerte aus ($z = \frac{x - \bar{x}}{s}$) und be-

stimmen über das Intervall, in das der z-Wert eines Schülers hineinfällt, die Note, die der Schüler erhält. Ein Schüler mit einem z-Wert von z. B. $+0,45$ würde nach dem Modell in Abbildung 36 die Note 2 erhalten.

- b) Sie gehen von den Prozenträngen der Rohwertverteilung aus ($PR =$

$$\frac{\text{cum}_i}{N} \cdot 100$$

) und bestimmen über die Prozentrangintervalle die Noten

für die Schüler. So würde z. B. ein Schüler mit dem Prozentrang von $PR = 40,6$ nach dem Modell in Abbildung 36 die Note 3 erhalten.

Es ist nicht gleich, ob nach Verfahren a oder b die Noten bestimmt werden. Bestimmen Sie die Noten über die Prozentränge (Verfahren b), nehmen Sie auch gleichzeitig eine Normalisierung der Notenverteilung vor: die Umsetzung der Rohwerte in Prozentränge normalisiert die Verteilung der Rohwerte (vgl. Kap. 5.6). Werden dann die Noten den Prozentrangintervallen zugeordnet, so wird auch die Notenverteilung normal verteilt.

Werden hingegen die Noten den z-Werten der Rohwertverteilung zugeordnet (Verfahren a), sind die Noten so verteilt, wie die Rohwerte verteilt sind. Bei der Transformation der Rohwerte in z-Werte bleibt die

Form der Rohwertverteilung erhalten. Die Rohwerttransformation in z-Werte ist, wie schon an anderer Stelle gezeigt worden ist, linear. Die Transformation in Prozentränge ist nicht linear.

Es dürfte sich empfehlen, die Notenbestimmung immer linear, d. h. nach Verfahren a vorzunehmen. Das Leistungsbild der Schüler findet sich dann in den Noten auch so wieder, wie es sich in der Rohwertverteilung zeigt. Über die Prozentränge hingegen entsteht eine Notenverteilung, die fälschlich auf ein normalverteiltes Leistungsbild unter den Schülern schließen läßt.

Auf einen fest vorgegebenen Notenschlüssel (Tabelle 36) haben sich die Industrie- und Handelskammern für die von ihnen durchzuführenden Berufsprüfungen geeinigt⁴⁸.

Rohwert (Punkte)	Note
92 - 100	1
81 - 91	2
67 - 80	3
50 - 66	4
30 - 49	5
0 - 29	6

Tab. 36:

100-Punkte-Notenschlüssel der Industrie- und Handelskammern.

Der 100-Punkte-Schlüssel ist auf den jeweils in einem Test verwendeten Punktebereich umzusetzen. Für den Test Vitamine und Mineralstoffe ergibt sich bei maximal 19 Rohwerten z. B. ein Rohwertebereich für die Note 2 von $0,81 \cdot 19 \approx 15$ bis $0,91 \cdot 19 \approx 17$. Entsprechend sind die Bereiche für die anderen Noten zu berechnen. – Der 100-Punkte-Schlüssel ist ein strenger Maßstab. Erst bei 50% der maximal zu erreichenden Punkte kann die Note 4 vergeben werden. Im Beispielttest Vitamine und Mineralstoffe ergeben 50% von 19 Rohwerten abgerundet 9 Rohwerte. Es wird von 19 Rohwerten ausgegangen, da Aufg. 20 aufgrund negativer Trennschärfe nicht berücksichtigt wird, vgl. S. 141.

13.2 Problematik eines Zensierungsmodells nach der Normalverteilung

Die an der Normalverteilung ausgerichteten Notenverteilungsmodelle sind problematisch. Bei allen Modellen steht von vornherein fest, wie viele Schüler eine mangelhafte oder eine gute Leistung bescheinigt bekommen. Nach dem Modell in Abbildung 35 sind es z. B. 7%, nach dem Modell in Abbildung 36 10% der Schüler, welche die Note 5 erhalten werden. Dies ist unabhängig davon, wie gut ihre Leistung absolut ist. Umgekehrt, wenn alle nur ein Minimum an Lernfortschritt zeigen, wird ein

bestimmter Prozentsatz der Schüler die Note 1 erreichen, unabhängig von ihrer – absolut gesehenen – geringen Leistung.

Bei der Notenbestimmung, soll sie anhand der hier vorgelegten Modelle vorgenommen werden, muß auch die vom Lehrer sachlich zu begründende Leistungsanforderung einfließen. Die Rohwerte werden nach Verfahren a über z-Werte gemäß einem Verteilungsmodell in Noten umgesetzt. Der Lehrer hat dann zu entscheiden, ob die dabei ermittelten Noten mit den Leistungen korrespondieren, die von seinen Leistungsanforderungen her gerecht erscheinen. Er wird aufgrund pädagogischer Entscheidungen, die über den statistisch-rechnerischen Entscheidungen stehen, in die Notenverteilung eingreifen.

Am Beispiel des Testes Vitamine und Mineralstoffe wird in Tabelle 37 der Vorgang der Notenbestimmung demonstriert. Den z-Werten der Rohwerte dieses Testes lassen sich nach einem Notenverteilungsschlüssel – hier wurde der in Abbildung 35 gewählt – rasch Noten zuordnen. So entfällt z. B. die Note 3 auf den z-Wertebereich von $-0,5$ bis $+0,5$. Dies entspricht den Rohwerten 11 bis einschließlich 13. Mit anderen Worten, die Schüler, die insgesamt 11, 12 oder 13 Rohwerte erreichten, erhalten für diese Testleistung die Note 3.

Rohwert	z-Wert	Häufigkeit	Note	Häufigkeit
18	2,0	1	1	2
17	1,7	1		
16	1,3	2	2	7
15	1,0	2		
14	0,7	3		
13	0,3	5	3	12
12	0,0	2		
11	-0,3	5		
10	-0,7	3	4	7
9	-1,0	0		
8	-1,3	4		
7	-1,7	1	5	2
6	-2,0	1		
Summe		30		30

Tab. 37:

Umsetzung der Rohwerte des Testes Vitamine und Mineralstoffe in Noten nach dem Verteilungsschema in Abbildung 35.

Wird wie oben die Notenverteilung an der Normalverteilung ausgerichtet, fällt der Schnitt für eine nicht mehr ausreichende Leistung bei 7 und weniger Rohwerten. Eine Leistung, die für gut gehalten wird, setzt insgesamt 14 Rohwerte ein. Diese und die anderen Einschnitte wurden statistisch-rechnerisch ermittelt. Sie haben nun zu entscheiden, ob die über diese Einschnitte ermittelten Noten mit Leistungen korrespondie-

ren, die Ihren Leistungsanforderungen entsprechen. Trifft dies nicht zu, müßten Sie korrigierend in die Notenverteilung eingreifen. So ließe sich der Einschnitt für die Note 2 über die Rohwertzahl 14 legen, wenn Sie der Auffassung sind, daß z. B. bei diesem Test die Note 2 erst ab 15 insgesamt erreichten Rohwerten vergeben werden kann. Die pädagogische Entscheidung führt zu der Note und nicht die statistisch-rechnerische Entscheidung.

14. Aktualisierung statistischer Berechnungen mit moderner Software

Thomas Heinrich

In diesem Kapitel sollen mit Hilfe des weit verbreiteten Tabellenkalkulationsprogrammes Microsoft EXCEL für Windows einige der in Teil A aufgeführten Abbildungen, Tabellen und die damit verbundenen Rechenoperationen in einfacher Weise dargestellt und deren Erarbeitung in EXCEL 7.0 kurz beschrieben werden. Zum leichteren Vergleich zwischen herkömmlicher Methode und Anwendung einer geeigneten Software werden zusätzlich Tabellen- bzw. Abbildungsnummern und die Seite, auf welcher die ursprünglichen Abbildungen, Tabellen und Rechenoperationen zu finden sind, angegeben.

Vor allem der Computerlaie soll dadurch in der Lage sein, die aufgeführten Beispiele anhand der Beschreibung mit Hilfe des arbeitserleichternden Tabellenkalkulationsprogrammes nachvollziehen und nötigenfalls selbständig ausüben zu können. Mit Hilfe einiger kurzer Hinweise bezüglich der Software muß der Leser notwendigerweise die erforderlichen Arbeitsschritte im Handbuch für EXCEL 7.0 nachlesen.

14.1 Urliste, Rangfolgenverteilung und Primäre Häufigkeitsverteilung (f)

Die Erstellung der Urliste (Definition 1, Teil A, S. 14) für den Test Vitamine und Mineralstoffe erfolgt mit EXCEL relativ einfach. Dazu werden in eine Spalte (z.B. Spalte A) die Namen aller beteiligten Schüler und in die Nachbarspalte (z.B. Spalte B) die erzielten Rohwerte der jeweiligen Schüler eingetragen. Als Spaltenüberschrift empfiehlt es sich zum einen „Rohwertträger Schüler“ und zum anderen „Rohwert (RW)“ zu verwenden. Bei Formatunterschieden (fett, kursiv, Schriftgröße, o.ä.) zwischen Spaltenüberschrift und eigentlichen Daten wird die Überschrift als solche erkannt und von den eingetragenen Daten innerhalb der Software abgegrenzt.

Die Rangfolge der Rohwerte bleibt zunächst noch unberücksichtigt, so daß das Ergebnis vorerst wie in Tabelle 1 aussehen könnte.

Um aus dieser einfachen Tabelle eine Rangfolgenverteilung mit absteigender Anordnung der Rohwerte im Test Vitamine und Mineralstoffe (vgl. Tabelle 1, Teil A, S. 15) zu erhalten, hilft der Befehl SORTIEREN im Fenster DATEN weiter. Damit EXCEL erkennt, welche Daten sortiert werden sollen, muß zunächst mindestens eine Zelle innerhalb der zu sortierenden Auflistung markiert werden. Der Befehl SORTIEREN ist im zu öffnenden Fenster DATEN in der Menüleiste zu finden. Als Sortierargument werden unter anderem die einzelnen Spaltenüberschriften vorgegeben. In diesem Fall muß SORTIEREN nach ROHWERT (RW) und ABSTEIGEND ausgewählt werden. Der größte Roh-

Rohwertträger Schüler	Rohwert (RW)	Rohwertträger Schüler	Rohwert (RW)
Thomas B.	10	Gudrun P.	8
Marlies G.	6	Helga W.	14
Ute B.	12	Friedel S.	17
Sonja S.	10	Michael E.	8
Ilse M.	8	Jutta H.	11
Britta A.	13	Christine W.	14
Marion B.	11	Christa W.	14
Erika B.	11	Cornelia F.	11
Angela R.	13	Elvira G.	16
Siegrid L.	11	Bärbel W.	8
Carla G.	15	Christa H.	18
Hilde M.	7	Astrid P.	12
Regina N.	13	Christel G.	13
Ingeborg E.	10	Beatrix H.	15
Karin E.	13	Ilona W.	16

Tabelle 1:

Verteilung der Rohwerte im Test Vitamine und Mineralstoffe, unsortiert

wert soll entsprechend dem besten Schüler an erster Stelle stehen. Nach Beendigung des Sortiervorganges stellt sich die nach Rohwerten sortierte Liste wie in Tabelle 2 dar.

Rohwertträger Schüler	Rohwert (RW)	Rohwertträger Schüler	Rohwert (RW)
Christa H.	18	Astrid P.	12
Friedel S.	17	Marion B.	11
Elvira G.	16	Erika B.	11
Ilona W.	16	Siegrid L.	11
Carla G.	15	Jutta H.	11
Beatrix H.	15	Cornelia F.	11
Helga W.	14	Thomas B.	10
Christa W.	14	Sonja S.	10
Christine W.	14	Ingeborg E.	10
Britta A.	13	Ilse M.	8
Angela R.	13	Gudrun P.	8
Regina N.	13	Michael E.	8
Karin E.	13	Bärbel W.	8
Christel G.	13	Hilde M.	7
Ute B.	12	Marlies G.	6

Tabelle 2:

Verteilung der Rohwerte im Test Vitamine und Mineralstoffe, absteigend sortiert

In EXCEL steht die Funktion RANG, die eine Aussage über die relative Position eines Rohwertträgers innerhalb eines getesteten Personenkreises macht, zur Verfügung und kann in einer weiteren Spalte der oben abgebildeten Tabelle eingefügt werden. In dem unter der Menüleiste EINFÜGEN sich öffnenden Fenster wird der Befehl FUNKTION ausgewählt. Es erscheint die erste Seite des Funktions-Assistenten. Sämtliche bereits im Programm vorgegebene Funk-

tionen, die unter verschiedenen Kategorien alphabetisch angeordnet sind, werden aufgelistet. Nach Auswahl einer Funktion benötigt der Rechner weitere Daten, die auf der zweiten Seite des Funktions-Assistenten (WEITER>) erfragt werden. Sind alle erforderlichen Angaben eingetragen worden, wird der Wert, welcher anschließend in der entsprechenden Zelle erscheint, im zweiten Fenster des Funktions-Assistenten oben rechts angegeben. Eine Übernahme in die ausgewählte Zelle erfolgt durch Beendigung des Funktions-Assistenten. Da die Formel RANG in alle Zellen der Spalte „Rang“ eingetragen werden muß, empfiehlt es sich bei der Erstellung auf relative und absolute Zellbezüge zu achten. Ausführliche Erläuterungen über relative und absolute Zellbezüge sind im Handbuch für EXCEL nachzulesen.

Das Ergebnis der erweiterten Tabelle findet man in Tabelle 3.

Rohwertträger Schüler	Rohwert	Rang	Rohwertträger Schüler	Rohwert	Rang
Christa H.	18	1	Astrid P.	12	15
Friedel S.	17	2	Marion B.	11	17
Elvira G.	16	3	Erika B.	11	17
Ilona W.	16	3	Siegrid L.	11	17
Carla G.	15	5	Jutta H.	11	17
Beatrix H.	15	5	Cornelia F.	11	17
Helga W.	14	7	Thomas B.	10	22
Christa W.	14	7	Sonja S.	10	22
Christine W.	14	7	Ingeborg E.	10	22
Britta A.	13	10	Ilse M.	8	25
Angela R.	13	10	Gudrun P.	8	25
Regina N.	13	10	Michael E.	8	25
Karin E.	13	10	Bärbel W.	8	25
Christel G.	13	10	Hilde M.	7	29
Ute B.	12	15	Marlies G.	6	30

Tabelle 3:

Rangfolgenverteilung der Rohwerte im Test Vitamine und Mineralstoffe, N = 30 Schüler, maximal erreichbarer Rohwert 20, vgl. Tabelle 1, Teil A, S.15

Beim Vergleich zwischen der Tabelle 1 im Teil A auf Seite 15 und der in EXCEL erstellten Tabelle fällt zunächst auf, daß der Rang in EXCEL stets ganzzahlig ist. EXCEL vergibt bei gleichen Rohwerten den betreffenden Schülern die gleiche Rangzahl. Dabei wird jedoch nicht das Mittel der betreffenden Rangzahlen der Schüler mit gleichen Rohwerten errechnet, sondern die der vorausgehenden Rangzahl folgende vergeben (vgl. Beispiel, Teil A, S. 15).

Die Grundausswertung des Tests Vitamine und Mineralstoffe (Tabelle 2, Teil A, S. 16) kann in EXCEL ebenso vereinfacht bearbeitet werden. Tabelle 1 dient dazu als Grundlage und erspart einige Schreibarbeit, da sämtliche Daten bis auf Schülernamen gelöscht und durch Aussagen über Bearbeitung der einzelnen Testaufgaben (+ = Aufgabe richtig, 0 = Aufgabe falsch, / = Aufgabe nicht beantwortet) ergänzt werden können. Die Ermittlung der Punktwerte des einzelnen Schülers und die Anzahl, wie viele Schüler eine bestimmte Aufgabe richtig gelöst haben, erfolgt über die Funktion ZÄHLENWENN, mit Hilfe

derer in diesem Fall die „+“-Zeichen (entspricht den gelösten Aufgaben) des jeweiligen Schülers bzw. der jeweiligen Aufgabe gezählt werden. Diese Funktion ist im Funktions-Assistenten unter der Kategorie „Mathematik & Trigonometrie“ (oder „Alle“) zu finden. Da die Formel wiederum in mehrere Zellen kopiert werden muß, sollte auf relative und absolute Zellbezüge geachtet werden.

Die in EXCEL erstellte Grundausswertung des Tests Vitamine und Mineralstoffe wird in Tabelle 4 gezeigt.

Rohwertträger Schüler	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Punktwert Schüler
Christa H.	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0	+	+	+	/	18
Friedel S.	0	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0	+	+	+	/	17
Ilona W.	0	+	+	+	+	+	+	+	+	+	+	+	+	+	0	0	+	+	+	/	16
Elvira G.	+	+	+	+	+	+	0	+	0	0	+	+	+	+	+	0	+	+	+	+	16
Carla G.	+	0	+	+	+	0	+	+	0	+	+	+	+	+	+	0	+	+	+	/	15
Beatrix H.	+	0	+	+	0	+	+	+	+	+	+	+	+	+	/	+	0	+	+	/	15
Christine W.	0	+	+	+	0	+	0	+	0	+	+	+	+	+	0	+	0	+	+	/	14
Christa W.	+	+	+	+	0	+	0	+	/	+	0	+	+	+	+	0	+	+	+	0	14
Helga W.	+	+	0	+	0	+	+	+	0	+	0	+	+	+	+	0	+	+	0	+	14
Angela R.	+	0	+	+	0	0	+	+	+	+	+	0	+	+	0	+	/	+	+	/	13
Christel G.	+	0	+	+	0	0	+	+	+	+	+	+	+	+	+	0	+	+	+	/	13
Regina N.	0	0	+	+	+	0	+	+	0	+	+	+	+	+	0	0	+	+	+	/	13
Britta A.	+	+	+	+	0	+	+	+	0	0	+	+	0	0	0	0	+	+	+	/	13
Karin E.	0	+	+	+	+	0	0	+	+	0	0	+	+	+	+	0	+	+	+	+	13
Astrid P.	+	0	+	+	+	+	0	+	0	+	+	/	+	+	0	0	0	+	+	/	12
Ute B.	+	0	+	+	/	0	+	+	0	+	+	/	+	+	0	+	0	+	+	/	12
Erika B.	+	0	0	+	+	0	+	+	+	+	+	0	/	0	0	0	0	+	+	/	11
Marion B.	0	+	+	+	0	+	+	0	/	0	+	+	+	+	0	0	0	+	0	+	11
Jutta H.	0	0	+	+	+	0	+	+	/	0	/	+	+	+	+	0	0	+	/	+	11
Cornelia F.	0	0	+	+	0	0	+	+	0	0	/	+	+	+	+	0	0	+	+	/	11
Siegrid L.	+	0	+	+	0	+	+	+	0	/	+	0	+	+	+	0	0	+	0	/	11
Sonja S.	+	0	+	+	/	+	/	+	+	+	+	/	+	0	0	0	/	/	+	/	10
Thomas B.	0	0	/	+	+	0	+	0	0	+	+	+	+	0	/	+	0	+	+	/	10
Ingeborg E.	0	+	+	+	+	+	0	+	0	/	0	+	0	0	0	0	0	+	+	+	10
Gudrun P.	+	0	/	+	+	+	/	0	0	+	+	/	/	/	/	0	0	+	+	/	8
Michael E.	0	0	+	+	+	0	0	+	0	+	0	/	+	+	/	0	0	+	/	/	8
Bärbel W.	0	0	+	+	0	+	+	0	0	/	+	0	0	/	0	0	+	+	/	/	8
Ilse N.	+	0	+	+	+	0	0	0	0	/	+	0	0	/	0	0	+	+	+	+	8
Hilde M.	0	0	+	+	0	+	+	0	0	0	+	+	0	/	0	0	0	+	0	/	7
Marlies G.	0	0	+	/	0	+	0	+	0	0	0	+	0	0	0	0	0	0	+	+	6
Aufgabe gelöst	16	11	26	29	16	17	20	25	10	18	18	23	23	20	9	6	9	28	24	10	

Tabelle 4:

Grundausswertung des Tests Vitamine und Mineralstoffe, vgl. Tabelle 2, Teil A, S. 16

Die Anordnung der Rohwerte aus dem Test Vitamine und Mineralstoffe in einer Häufigkeitsverteilung läßt sich ebenfalls mit der Funktion ZÄHLEN-WENN realisieren. Die vorhandene Tabelle 2 kann durch zwei weitere Spalten „Rohwert (RW)“ und „Häufigkeit (f)“ erweitert werden.

In die Spalte „Rohwert (RW)“ werden alle Rohwerte, vom größten bis zum kleinsten auftretenden, in eine neue Spalte einmal eingetragen. In der Nachbarspalte wird über die Funktion ZÄHLENWENN ermittelt, wie oft ein bestimmter Rohwert in der Spalte „Punktwert Schüler“ auftritt. Beim Kopieren der Formel in die restlichen Zellen dieser Spalte muß sich das Suchkriterium von Rohwert zu Rohwert (Zeile zu Zeile) verändern: 18, 17, 16, ... 6.

Zur Kontrolle, ob alle teilgenommenen Schüler in der Häufigkeitsspalte berücksichtigt wurden, empfiehlt es sich am Ende dieser Spalte die Summe aller Häufigkeiten zu bilden, die mit der Schülerzahl (N=30) identisch sein muß. Die Funktion SUMME kann in EXCEL ebenfalls über den Funktions-Assistenten aufgerufen werden.

Als vorläufiges Ergebnis erhält man die Tabelle 5.

Rohwert (RW)	Häufigkeit (f)
18	1
17	1
16	2
15	2
14	3
13	5
12	2
11	5
10	3
9	0
8	4
7	1
6	1
Summe	30

Tabelle 5:

Primäre Häufigkeitsverteilung der erreichten Rohwerte im Test Vitamine und Mineralstoffe, maximal erreichbarer Rohwert 20, vgl. Tabelle 3, Teil A, S. 17

Die Darstellung von Tabelle 4, Teil A, S. 18, „Rangfolgenverteilung und Primäre Häufigkeitsverteilung der Rohwerte in einem ‘Lesereifetest’“ erfordert keine Anwendung neuer Funktionen. Sie kann zur Wiederholung der beschriebenen Vorgehensweise der Tabellen 1 bis 3 und zur Einarbeitung in die Software dienen. Außerdem werden in den noch ausstehenden Kapiteln auch Berechnungen bezüglich dieses ‘Lesereifetests’ aufgeführt. Eine mögliche Aufbereitung von Tabelle 4, Teil A, S. 18, mit Hilfe von EXCEL zeigt hier Tabelle 6.

Rohwert	Rang	Rohwert	Häufigkeit	Rohwert	Häufigkeit
112	1	112	1	76	0
109	2	111	0	75	3
106	3	110	0	74	1
105	4	109	1	73	0
104	5	108	0	72	1
100	6	107	0	71	1
97	7	106	1	70	1
95	8	105	1	69	1
95	8	104	1	68	1
93	10	103	0	67	0
91	11	102	0	66	1
90	12	101	0	65	0
89	13	100	1	64	0
84	14	99	0	63	0
84	14	98	0	62	1
83	16	97	1	61	0
82	17	96	0	60	0
81	18	95	2	59	1
80	19	94	0	58	1
78	20	93	1	57	0
75	21	92	0	56	0
75	21	91	1	55	0
75	21	90	1	54	0
74	24	89	1	53	0
72	25	88	0	52	0
71	26	87	0	51	1
70	27	86	0	50	0
69	28	85	0	49	0
68	29	84	2	48	0
66	30	83	1	47	1
62	31	82	1	46	0
59	32	81	1	45	0
58	33	80	1	44	1
51	34	79	0	Summe N 36	
47	35	78	1		
44	36	77	0		

Tabelle 6:

Rangfolgenverteilung und Primäre Häufigkeitsverteilung der Rohwerte in einem 'Lesereifetest', vgl. Tabelle 4, Teil A, S. 18

14.2 Median (M_d), Modalwert (M_o), Arithmetischer Mittelwert ($\bar{x}$)

Die Berechnung des Medians kann mit Hilfe der bereits vorliegenden Funktion in der Kategorie STATISTIK im Funktions-Assistenten für Tabelle 6 ('Lesereifetest') relativ leicht durchgeführt werden. Die Funktion MEDIAN benötigt als Ausgangsdaten alle einzeln aufgelisteten Rohwerte. Das Ergebnis des Medians für den 'Lesereifetest' muß nach Beenden des Funktions-Assistenten 80,5 lauten.

Beim Testbeispiel Vitamine und Mineralstoffe in Tabelle 2 ermittelt sich der Median scheinbar genauso einfach über den Gesamtbereich der Punktwerte der einzelnen Schüler wie beim 'Lesereifetest' in Tabelle 6. Da bei diesem Test jedoch nahe des Medians gleiche Rohwerte mehrfach auftreten, ist mit der einfachen Formel nach EXCEL, die vermutlich diese Tatsache nicht berücksichtigt, der Median nicht mehr korrekt zu bestimmen. Die für diesen Fall in sechs Schritten untergliederte Berechnung des Medians läßt sich mit EXCEL nicht ohne weiteres nachvollziehen.

Die zur schrittweisen Ermittlung des Medians notwendige Tabelle, die die Spalten „Rohwert (RW)“, „Häufigkeit (f)“ und „Cumulierte Häufigkeit cum_i“ umfaßt, ist durch Erweiterung von Tabelle 5 ohne großen Aufwand zu gewinnen. Ihr fehlt nur die Spalte „Cumulierte Häufigkeit cum_i“.

Nach Definition 5, Teil A, S. 21, gibt die Cumulierte Häufigkeit die Zahl der Rohwertträger wieder, die unter einem gegebenen Rohwert + 0,5 liegen. Mit Hilfe der Spalte „Häufigkeit (f)“ in Tabelle 5 und der Funktion SUMME in EXCEL läßt sich die Zahl der Rohwertträger bis zu einem gewünschten Rohwert ermitteln. Beim Kopieren der erstellten Formel für eine Zelle in alle restlichen der Spalte „Cumulierte Häufigkeit cum_i“ erweist sich als Vorteil auf relative und absolute Zellbezüge zu achten, um unnötige Mehrarbeit einzusparen.

Als Ergebnis erhält man Tabelle 7, die zur Bestimmung des Medians in Schritten notwendig ist.

Rohwert (RW)	Häufigkeit (f)	Cumulierte Häufigkeit cum _i
18	1	30
17	1	29
16	2	28
15	2	26
14	3	24
13	5	21
12	2	16
11	5	14
10	3	9
9	0	6
8	4	6
7	1	2
6	1	1
Summe	30	

Tabelle 7:

Bestimmung des Medians für die Ergebnisse im Test Vitamine und Mineralstoffe, vgl. Tabelle 5, Teil A, S. 22

Berechnet man den Median des Tests Vitamine und Mineralstoffe mittels Funktion analog zum 'Lesereifetest', erhält man wie nach Berechnung des Medians in Schritten als Ergebnis 12. Daß zwischen Median-Berechnung in EXCEL und nach Schritten ein Unterschied besteht, wird erst im Ergebnis für

das fiktive Beispiel in Tabelle 6, Teil A, S. 23, deutlich. Der nach Schritten berechnete Median lautet hier 3,125 (Median nach EXCEL = 3).

Die Erstellung der Tabelle für das Beispiel in Tabelle 6, Teil A, S. 23, zur Berechnung des Medians nach EXCEL wird nicht mehr ausführlich beschrieben. Sie kann als Übungsbeispiel dienen und wird zur Kontrolle nachfolgend aufgeführt.

Rohwert (RW)	Häufigkeit (f)	cum, f	RW
5	3	15	5
4	3	12	5
3	4	9	5
2	3	5	4
1	2	2	4
Median			3
			3
			3
			3
			3
			2
			2
			2
			1
			1

Tabelle 8 :

Fiktives Beispiel zur Berechnung des Medians, vgl. Tabelle 6, Teil A, S. 23

Übung 2 auf Seite 23 in Teil A wird hier bewußt nicht behandelt, da der Leser anhand der Berechnung nach Schritten den Wert des Medians exakt ermitteln soll. Auch bei dieser Übung ließe sich ein Unterschied zwischen beiden Berechnungsarten (Median nach Schritten: Md = 42,9, Median nach EXCEL: Md = 43) von 0,1 feststellen.

Bei der Ermittlung des Modalwertes aus den Rohwerten des 'Lesereifetests' kann ebenfalls mit dem Funktions-Assistenten gearbeitet werden. Der Modalwert ist als Funktion in EXCEL vorgegeben und benötigt als Daten jeden erzielten Rohwert der einzelnen Testpersonen. Er beträgt im 'Lesereifetest' 75.

Bei der Berechnung des Modalwertes für den Test Vitamine und Mineralstoffe ermittelt EXCEL nur den Wert 13, obwohl der Rohwert 11 ebenfalls fünfmal auftritt. Das Problem besteht offensichtlich darin, daß nur der zuerst am häufigsten vorkommende Rohwert im Programm als Modalwert zählt (hier: 13, und nicht 11, da die Tabelle nach Rohwerten absteigend sortiert ist.). Sonderfälle, wie sie auf S. 25 im Teil A beschrieben werden, bleiben demnach unberücksichtigt.

In der graphischen Darstellung der bimodalen Häufigkeitsverteilung im Testbeispiel Vitamine und Mineralstoffe (vgl. Abbildung 5, Teil A, S. 31) ist auf den ersten Blick ersichtlich, daß in dieser Testreihe zwei Modalwerte auftreten. Da auch in EXCEL die Möglichkeit besteht, ausgewählte Daten in einem

Diagramm graphisch darzustellen, ergibt sich zwischen Berechnung des Modalwertes und Graphik aus oben angegebenem Grund ein Widerspruch. Werden die Daten „Rohwert (RW)“ und „Häufigkeit (f)“ aus Tabelle 7 in einem Koordinatensystem mit den Rohwerten auf der Rechtswertachse und der Häufigkeit der entsprechenden Rohwerte auf der Hochwertachse eingetragen, wird anhand des Graphen deutlich, daß in dieser Testauswertung zwei Modalwerte bei den Rohwerten 11 und 13 existieren. Diese beiden Rohwerte wurden im Test jeweils fünfmal erzielt.

Um ein Diagramm in EXCEL zu erstellen, muß in der Menüleiste unter EINFÜGEN der Befehl DIAGRAMM aktiviert werden. Nach der ersten Auswahl, in welcher Arbeitsmappe das Diagramm erstellt werden soll, müssen folgende Schritte des Diagramm-Assistenten wie gewünscht beantwortet werden:

1. Schritt: Auswahl / Markierung des darzustellenden Bereiches
2. Schritt: Auswahl des Diagrammtyps
3. Schritt: AutoFormat für den ausgesuchten Diagrammtyp
4. Schritt: Ausrichtung der Datenreihen auf entsprechende Achsen, Legende
5. Schritt: Diagrammtitel, Achsenbeschriftung

Beide den Schritten 4 und 5 wird bereits das Beispieldiagramm, das die eingegebenen Daten berücksichtigt, angezeigt. Änderungen innerhalb des Diagramm-Assistenten können jederzeit durch Vor- und Zurückblättern zu den verschiedenen Schritten verwirklicht werden. Diese übertragen sich unmittelbar auf das Beispieldiagramm. Nach Beendigung des Diagramm-Assistenten erscheint das erstellte Diagramm in der ausgewählten Arbeitsmappe. Durch Doppelklicken auf die Graphik kann diese auch noch im Nachhinein bearbeitet werden (siehe unter EINFÜGEN bei ausgewählter Graphik). Eine sehr ausführliche Beschreibung, wie Diagramme erstellt und bearbeitet werden können, ist im Handbuch abgedruckt.

Abbildung 1 zeigt eine Möglichkeit, die Daten „Rohwert (RW)“ und „Häufigkeit (f)“ aus Tabelle 7 in eine Graphik zu übertragen.

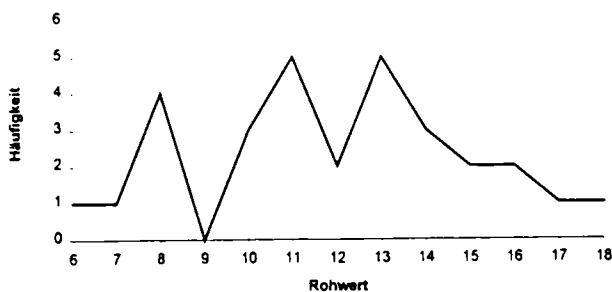


Abbildung 1:
Häufigkeitsverteilung im Testbeispiel Vitamine und Mineralstoffe, Daten aus Tabelle 7, vgl. Abbildung 5, Teil A, S. 31

Um das arithmetische Mittel $\bar{x}$ zu bilden, muß die Summe der Rohwerte durch die Anzahl der Rohwertträger dividiert werden. Mühsames Ein- und möglicherweise Vertippen auf dem Taschenrechner entfällt, da auch diese Berechnung in EXCEL mit der Funktion MITTELWERT in einem Schritt durchgeführt werden kann. Als Ergebnis des arithmetischen Mittels $\bar{x}$ berechnet EXCEL für den Test Vitamine und Mineralstoffe den Wert 11,93333333.

Da für die Aussagekraft des Mittels zwei Nachkommastellen völlig ausreichend sind, kann das Ergebnis mit Hilfe der Funktion RUNDEN auf die gewünschte Anzahl der Nachkommastellen begrenzt werden. Um den Mittelwert auf zwei Stellen nach dem Komma zu runden, muß eine verschachtelte Funktion erstellt werden. Die Vorgehensweise bei der Erstellung einer solchen verschachtelten Funktion wird im Handbuch genau erklärt.

Der Mittelwert des 'Lesereifetests' in Tabelle 4 kann ebenso nach oben angegebenem Schema über die Funktionen RUNDEN und MITTELWERT gewonnen werden. Das Ergebnis lautet auf zwei Stellen gerundet 80,39.

14.3 Variationsweite (V), Quartilabweichung (Q), Standardabweichung (s), Varianz (s²)

Die als Variationsweite definierte Differenz zwischen dem größten und dem kleinsten Wert in einer Rohwertverteilung läßt sich durch die Funktionen MAX und MIN, die beide im Funktions-Assistenten vorgegeben sind, ermitteln. Diese Funktionen suchen innerhalb sämtlicher Rohwerte im Test Vitamine und Mineralstoffe nach dem Maximal- (RW = 18) bzw. Minimalwert (RW = 6). Werden beide Funktionen über das Minuszeichen verknüpft, erhält man als Ergebnis die Variationsweite $V = 12$.

Die Quartilabweichung ist als Funktion in EXCEL nicht direkt vorgegeben. Mit der Funktion QUARTILE im Funktions-Assistenten kann man jedoch

- über Quartil gleich 0 den kleinsten Wert einer Matrix,
- über Quartil gleich 1 das untere Quartil (Q_1 bzw. P_{25}),
- über Quartil gleich 2 den Median (Q_2 bzw. P_{50}),
- über Quartil gleich 3 das obere Quartil (Q_3 bzw. P_{75}) und
- über Quartil gleich 4 den größten Wert einer Matrix

ermitteln. Bei dieser Funktion muß dann die Matrix (hier: Rohwerte) und das gewünschte Quartil (0; 1; 2; 3; 4) angegeben werden.

Um die Quartilabweichung Q zu bilden, wird zunächst die Interquartilweite $Q_3 - Q_1$ durch Verknüpfung beider Quartile über das Minuszeichen errechnet. Diese Formel in Klammern gesetzt und durch die Zahl 2 geteilt (vgl. Formel, Teil A, S. 33) ergibt für die Quartilabweichung den Wert 2. In Tabelle 9 sind alle Ergebnisse der Quartile und der Quartilabweichung für die Rohwerte aus dem Test Vitamine und Mineralstoffe aufgeführt.

0. Quartil	6
1. Quartil (= Q ₁)	10
2. Quartil (= Q ₂)	12
3. Quartil (= Q ₃)	14
4. Quartil	18
Quartilabweichung	2

Tabelle 9:

Berechnung der Quartile und der Quartilabweichung in EXCEL für den Test Vitamine und Mineralstoffe, vgl. Teil A, S. 34

Da durch das 0.Quartil bzw. das 4.Quartil jeweils der größte bzw. kleinste Wert innerhalb einer Matrix ermittelt wird, hätte sich auf diese Weise auch die Variationsweite berechnen lassen können.

Die einprogrammierte Funktion STABWN - **nicht STABW**, da die Standardabweichung in diesem Fall nur anhand einer Stichprobe geschätzt werden würde - erleichtert die mühsame Arbeit bei der Berechnung der Standardabweichung mit dem Taschenrechner für den Test Vitamine und Mineralstoffe. Für die in einer Matrix eingegebenen Werte errechnet der PC per Tastendruck die Standardabweichung. Allerdings wird ein Ergebnis mit neun Stellen nach dem Komma angegeben, welches sinnvollerweise über die Funktion RUNDEN auf zwei Stellen nach dem Komma begrenzt werden sollte. Für den Test Vitamine und Mineralstoffe erhält man als Standardabweichung gerundet den Wert 3,00.

Bei der Berechnung der Varianz für den Test Vitamine und Mineralstoffe kann analog zur Standardabweichung vorgegangen werden. Die Formel unterscheidet sich nur durch die Funktion VARIANZEN. Die Matrix mit den Rohwerten bleibt unverändert. Das Ergebnis für den Test Vitamine und Mineralstoffe beträgt gerundet 9,00. Auch hier kann über VARIANZ die Varianz ausgehend von einer Stichprobe nur geschätzt werden.

14.4 Perzentile und Prozentränge

Eine programmierte Funktion zur Ermittlung des Perzentils P_p ist innerhalb der Software nicht vorhanden. Die in Teil A auf S. 44 beschriebene Vorgehensweise zur Bestimmung eines bestimmten Perzentils muß daher weiterhin schrittweise erfolgen. Die dafür notwendige Tabelle wurde bereits erstellt (vgl. hier Tabelle 7)

Die Ermittlung des Prozentranges zu einem erzielten Rohwert im Test Vitamine und Mineralstoffe nach der Formel

$$PR = \frac{cum_f}{N} * 100 [\%]$$

auf S. 45 im Teil A gelingt in EXCEL relativ einfach. An Tabelle 7 wird die Spalte „Prozentrang (PR) [%]“ angehängt. Die Ergebnisse der Cumulierten Häufigkeit (f) werden in der neuen Spalte weiterverwendet und bilden den Zähler des Bruches. Die Anzahl der Schüler $N = 30$ ist konstant und bildet den Nenner. Dieser Bruch wird in Klammern gesetzt und mit 100 multipliziert. Da im Ergebnis wieder bis zu neun Stellen nach dem Komma auftreten können, wird dieses von Anfang an über die Funktion RUNDEN auf zwei Nachkommastellen begrenzt. Tabelle 10 zeigt das Ergebnis, welches mit Tabelle 13 auf S. 46 in Teil A verglichen werden kann.

Rohwert (RW)	Häufigkeit (f)	Cumulierte Häufigkeit cum _r	Prozentrang (PR) [%]
18	1	30	100,00
17	1	29	96,67
16	2	28	93,33
15	2	26	86,67
14	3	24	80,00
13	5	21	70,00
12	2	16	53,33
11	5	14	46,67
10	3	9	30,00
9	0	6	20,00
8	4	6	20,00
7	1	2	6,67
6	1	1	3,33
Summe	30		

Tabelle 10:
Bestimmung der Prozentränge (PR) für die Rohwerte im Test Vitamine und Mineralstoffe, vgl. Tabelle 13, Teil A, S. 46

Um die für eine Zeile erstellte Formel auf alle anderen notwendigen Zellen innerhalb der Spalte „Prozentrang (PR) [%]“ zu übertragen, sollte bereits bei der Erstellung auf relative und absolute Zellbezüge geachtet werden.

14.5 Standardwerte z , Transformation der Standardwerte z

Um die Rohwerte des Tests Vitamine und Mineralstoffe gut interpretieren zu können, werden diese in sog. Standardwerte z umgewandelt. Dieser Standardwert zeigt an, wie viele Standardabweichungen ein bestimmter Rohwert einer Person über oder unter dem Mittelwert liegt. Da eine Funktion zur Berechnung der Standardwerte in EXCEL jedoch nicht vorgegeben ist, muß die Formel zur Ermittlung des Standardwertes

$$z = \frac{x - \bar{x}}{s}$$

nach Definition 15, Teil A, S. 55, selbst erstellt werden. Als Basis dient dabei die um die Spalten „Rohwert (RW)“ und „Häufigkeit (f)“ (Tabelle 5) erweiterte Tabelle 4. Innerhalb dieser Tabelle werden noch mit Hilfe der Funktionen MITTELWERT und STABWN, die bereits in den vorausgegangenen Kapiteln und Anwendung fanden, der arithmetische Mittelwert $\bar{x}$ und die Standardabweichung s berechnet. Auf ganze Zahlen gerundet wird in die Formel für den Standardwert z als Mittelwert $\bar{x} = 12$ und als Standardabweichung $s = 3$ eingesetzt. In einer neuen Spalte „z-Wert“ verknüpft man den Rohwert, für welchen der z-Wert bestimmt werden soll, mit dem Mittelwert über das Minuszeichen und setzt diese Differenz in Klammern, um sie durch die Standardabweichung dividieren zu können. Die Ergebnisse der berechneten z-Werte für den Test Vitamine und Mineralstoffe zeigt Tabelle 11.

Rohwert (RW)	Häufigkeit (f)	z-Wert
18	1	2,0
17	1	1,7
16	2	1,3
15	2	1,0
14	3	0,7
13	5	0,3
12	2	0,0
11	5	-0,3
10	3	-0,7
9	0	-1,0
8	4	-1,3
7	1	-1,7
6	1	-2,0
Summe	30	

Tabelle 11:

Umsetzung der Rohwerte des Tests Vitamine und Mineralstoffe in Standardwerte z , vgl. Tabelle 15, Teil A, S. 55

Um nicht mehr mit gebrochenen und negativen Standardwerten z arbeiten zu müssen, können diese mit Hilfe von EXCEL relativ leicht in andere Standardwerte transformiert werden. Insgesamt wurden in Kapitel 5.3 im Teil A bereits vier verschiedene Grundformeln zur Transformation der Standardwerte z in verschiedene Standardwerte vorgestellt:

$$\text{T-Skala: } T = 50 + 10 \cdot z$$

$$\text{IQ-Skala: } IQ = 100 + 15 \cdot z$$

$$\text{Centilskala: } C = 5 + 2 \cdot z$$

$$\text{Z-Skala: } Z = 100 + 10 \cdot z$$

Da IQ- und C-Skalen in der Regel den Intelligenztests vorbehalten sind, zeigt Tabelle 12 die Umrechnung der Standardwerte z in T- und Z-Werte. Um diese Tabelle zu erhalten, wird Tabelle 11 um die Spalten „T-Wert“ und „Z-

Wert“ erweitert. Die Formeln zur Umrechnung der Standardwerte müssen - wie bei der Ermittlung der z-Werte - neu erstellt werden, da sie nicht als Funktion in EXCEL vorgegeben sind.

Rohwert (RW)	Häufigkeit (f)	z-Wert	T-Wert	Z-Wert
18	1	2,0	70	120
17	1	1,7	67	117
16	2	1,3	63	113
15	2	1,0	60	110
14	3	0,7	57	107
13	5	0,3	53	103
12	2	0,0	50	100
11	5	-0,3	47	97
10	3	-0,7	43	93
9	0	-1,0	40	90
8	4	-1,3	37	87
7	1	-1,7	33	83
6	1	-2,0	30	80
Summe	30			

Tabelle 12:

Umsetzung der Rohwerte des Tests Vitamine und Mineralstoffe in T- und Z-Werte, vgl. Tabelle 16, Teil A, S. 60

14.6 Berechnung des Korrelationskoeffizienten r

Im Teil A sind zur Berechnung des Korrelationskoeffizienten r auf den Seiten 86 und 88 zwei Formeln angegeben:

$$(1) \quad r = \frac{\sum z_x \cdot z_y}{N}$$

$$(2) \quad r = \frac{N \cdot (\sum x \cdot y) - (\sum x) \cdot (\sum y)}{\sqrt{[N \cdot \sum x^2 - (\sum x)^2] \cdot [N \cdot \sum y^2 - (\sum y)^2]}} \quad (\text{Maschinenformel})$$

Bei Formel (1) müssen zuerst die z-Werte für die Rohwerte zweier Variablenätze berechnet werden, um dann den Korrelationskoeffizienten r ermitteln zu können. Im Gegensatz zur Maschinenformel ist dieses Rechenverfahren relativ einfach und formelmäßig leicht zu behalten. Die Berechnung der einzelnen z-Werte für jeden Test nimmt jedoch viel Zeit in Anspruch, da in EXCEL keine geeignete umfassende Funktion vorgegeben wird.

Der Korrelationskoeffizient r nach der Maschinenformel (Formel 2) lässt sich dagegen direkt aus den Rohwerten zweier Variablenätze berechnen. Diese Formel ist somit entsprechend umfangreicher. EXCEL verfügt über die Funk-

tion KORREL im Funktions-Assistenten, der den Korrelationskoeffizienten r ermittelt und nur die Rohwerte der Variablenätze benötigt. Die Funktion KORREL liefert den Korrelationskoeffizienten r einer zweidimensionalen Zufallsgröße, deren Werte in den jeweiligen Zellbereichen von zwei Matrizen stehen.

Im Beispiel aus Tabelle 25 im Teil A auf S. 87 haben sich 25 Schüler zwei zeitlich voneinander getrennten, verschiedenen Leistungstests unterzogen. Aus beiden Datensätzen soll der Korrelationskoeffizient r errechnet werden. Dazu müssen in einer Tabelle (hier: Tabelle 13) die Schüler und die erzielten Rohwerte in beiden Leistungstests eingetragen werden.

Schüler	Test I	Test II
	Rohwert (x)	Rohwert (y)
A	35	38
B	38	45
C	22	25
D	40	33
E	45	44
F	31	25
G	27	40
H	42	40
I	40	32
J	30	22
K	28	29
L	41	38
M	29	32
N	35	40
O	46	48
P	22	28
Q	21	30
R	33	26
S	37	39
T	41	42
U	26	30
V	22	31
W	33	22
X	42	40
Y	48	38

Tabelle 13:

Rohwerte der Leistungstests I und II, $N = 25$ Schüler, maximal erreichbarer Rohwert in jedem Test 50, vgl. Tabelle 25a, Teil A, S. 87

Über die Funktion KORREL erhält man ohne vorhergehende Berechnung des arithmetischen Mittelwertes $\bar{x}$ und der Standardabweichung s - beide Daten wären für die Formel (1) notwendig - den Korrelationskoeffizienten r . Er beträgt bei beiden Funktionen auf drei Stellen nach dem Komma gerundet 0,661.

Eine Berechnung des Korrelationskoeffizienten r mit Hilfe der Formel (1) über die z -Werte wäre in EXCEL problemlos durchführbar und könnte zur Kontrolle des ermittelten Wertes dienen. Diese Mehrarbeit bei der Berechnung der notwendigen z -Werte wird und sollte jedoch durch die bereits vorhandene Funktion KORREL erspart bleiben.

15. Neuere Entwicklungen

Für eine Testbeurteilung und Testerstellung sind eine Reihe von inhaltlichen Aktualisierungen anzusprechen, auf die hier in einem Erweiterungskapitel für die zweite Auflage nach 1980 kurz eingegangen werden soll. Dazu wird im folgenden dargestellt, inwieweit die hier zur Verdeutlichung von teststatistischen und testtheoretischen Begriffen herangezogenen Testverfahren noch bestehen bzw. welche Veränderungen sich ergeben haben bzw. sich ergeben. Im weiteren wird der Diskussionsstand zur klassischen Testtheorie, auf den das vorliegende Buch gründet, kurz angerissen. Damit soll allein eine erste Orientierung über neuere testtheoretische Ansätze gegeben werden. Die Ausführungen werden mit Hinweisen auf ausgewählte neuere Lehrbücher zur pädagogischen Diagnostik beschlossen.

Der hier angeführte Grundintelligenztest CFT3 nach R. Cattell, vgl. Teil A, Kap. 6.1., S. 73 ff., befindet sich seit 1971 unverändert in der Anwendung¹. Eine Neunormierung des Verfahrens ist in Vorbereitung.

Das Prüfsystem für Schul- und Bildungsberatung (PSB) von W. Horn, vgl. Teil A., Kap. 6.1, S. 76 ff., ist ebenso unverändert seit 1969 erhalten geblieben².

Die Drahtbiegeprobe (DBP) von G. A. Lienert, vgl. Teil A., Kap. 6.1, S. 79 ff., wird gleichfalls unverändert in zweiter, verbesserter Auflage seit 1967 vertrieben³.

Der im Teil B., Kap. 8, S. 99 erwähnte Intelligenz-Struktur-Test (IST) von R. Amthauer befindet sich in zweiter, erweiterter und verbesserter Auflage seit 1955 unverändert in Anwendung⁴. Eine erweiterte Ausgabe des IST stellt der ebenfalls von Amthauer erstellte Intelligenz-Struktur-Test (IST 70) dar, der in vierter, unveränderter Auflage seit 1973 lieferbar ist. Für diesen Test liegt heute auch eine computergestützte Fassung und ein Auswerteprogramm vor⁵.

Eine Übersicht über Tests gibt der hier zitierte Testkatalog des Berufsverbandes Deutscher Psychologen wieder. Es sei dabei an dieser Stelle noch darauf verwiesen, daß hinter dem oben genannten Prüfsystem für Schul- und Bildungsberatung (PSB) von W. Horn das Leistungsprüfsystem (LPS) vom gleichen Verfasser steht. Letzteres wird in zweiter, erweiterter und verbesserter Auflage seit 1983 zusammen mit einer computergestützten Fassung vertrieben. Die Eichung einiger Untertests wurde 1990 in Ost- und Westdeutschland überprüft⁶. Die hier im vorliegenden Lehr- und Studienbuch verwendeten Tests dienen allein der Verdeutlichung teststatistischer und testtheoretischer Kennwerte der klassischen Testtheorie. Für eine aktuelle Diskussion des diagnostischen Stellenwertes dieser Verfahren sei auf Kubinger verwiesen⁷.

Die Grundaussagen (Axiome) der klassischen Testtheorie können kritisiert werden, vgl. Teil B, Kap. 9.6, S. 122 f. Die psychologische Meßtheorie hat neue Modellansätze in Form einer probabilistischen Testtheorie entwickelt, eingeleitet durch das Rasch-Modell (1960). Eine einführende Übersicht gibt Plaum⁸, vertieft sei auf Kubinger verwiesen⁹. Bisher liegen allerdings nur we-

nige publizierte Tests vor, die diesen neueren testtheoretischen Modellansätzen folgen¹⁰. Klauer beklagt in diesem Zusammenhang, daß neuere Testmodelle entwickelt worden sind, „ohne daß ein nennenswerter Abruf von Seiten der Praxis erfolgt wäre“¹¹, und verdeutlicht dies besonders an der probabilistischen Testtheorie. Dagegen äußert sich Plaum dem Rasch-Modell gegenüber sehr kritisch. Darüber hinaus heißt es bei ihm: „Inwieweit sich die Intentionen der neueren testtheoretischen Ansätze tatsächlich in praktisch brauchbarer Weise realisieren lassen, ist ohne dies unter Diagnostikern umstritten“¹². Aus dem hier angeführten dürfte sich wohl ablesen lassen, daß die klassische Testtheorie ihren eigenen Stellenwert beibehält.

An neueren Lehrbüchern aus der pädagogischen Diagnostik seien hier abschließend allein drei Werke genannt. Dies ist zum einen eine neuere Auflage eines im vorliegenden Buch mehrfach herangezogenen amerikanischen Lehrbuches: *Educational and psychological measurement and evaluation* von Hopkins, Stanley und Hopkins in siebter, überarbeiteter Auflage von 1990¹³. Dieses Lehrbuch ist 1941 begründet worden und darf wohl als klassisch angesehen werden. Im deutschsprachigen Bereich hat sich das Lehrbuch der pädagogischen Diagnostik von Ingenkamp seit Mitte der achtziger Jahre etabliert und wohl einen ähnlichen Stellenwert wie das o. g. englischsprachige Werk erhalten¹⁴. Ergänzend zu nennen wäre als ein weiteres Lehrbuch die Diagnostik in pädagogischen Handlungsfeldern von Kleber¹⁵.

Zwischen dem Entwicklungsstand einer Testtheorie und Teststatistik und der Realisierung von Klassenarbeiten bzw. Schulaufgaben in der schulpraktischen Leistungsermittlung dürften auch heute noch in Deutschland große Brüche bestehen. Dies mag u. a. auch auf eine Test-Aversion des deutschen Intellektuellen zurückzuführen sein, wie Ingenkamp in einer Streitschrift anführt¹⁶. Dem kann nur entgegengewirkt werden, wenn ein auf Wissen gründender professioneller Umgang mit Tests gelernt wird. Dazu tragen die o. g. Lehrbücher bei. Das vorliegende Lehr- und Studienbuch soll auch hierzu einen Beitrag leisten.

Anmerkungen

¹ vgl. Testzentrale des Berufsverbandes Deutscher Psychologen (Hrsg.): Testkatalog 1996/97, Göttingen: Hogrefe 1996, S. 86

² vgl. Testzentrale des Berufsverbandes Deutscher Psychologen (Hrsg.): Testkatalog 1996/97, a. a. O., S. 60

³ vgl. Testzentrale des Berufsverbandes Deutscher Psychologen (Hrsg.): Testkatalog 1996/97, a. a. O., S. 279

⁴ vgl. Testzentrale des Berufsverbandes Deutscher Psychologen (Hrsg.): Testkatalog 1996/97, a. a. O., S. 77

⁵ vgl. Testzentrale des Berufsverbandes Deutscher Psychologen (Hrsg.): Testkatalog 1996/97, a. a. O., S. 77

- ⁶ vgl. Testzentrale des Berufsverbandes Deutscher Psychologen (Hrsg.): Testkatalog 1996/97, a. a. O., S. 79
- ⁷ vgl. K. D. Kubinger: Einführung in die psychologische Diagnostik, Weinheim: Psychologie-Verlags-Union 1995, S. 139 ff.
- ⁸ vgl. E. Plaum: Einführung in die Psychodiagnostik, Darmstadt: Primus-Verl. 1996, S. 69 ff.
- ⁹ vgl. K. D. Kubinger: Einführung in die psychologische Diagnostik, a. a. O., S. 64 ff.
- ¹⁰ vgl. Kubinger: Einführung in die psychologische Diagnostik, a. a. O., S. 69 f. und die dort angegebenen Tests.
- ¹¹ K. J. Klauer: Neuere Entwicklungen in der pädagogisch-psychologischen Diagnostik, in: Diagnostica 40 (1994)3, S. 244.
- ¹² E. Plaum: Einführung in die Psychodiagnostik, a. a. O., S. 82.
- ¹³ K. D. Hopkins, J. C. Stanley, R. B. Hopkins: Educational and Psychological Measurement and Evaluation, seventh edition, Englewood Cliffs, New Jersey: Prentice Hall 1990.
- ¹⁴ K. Ingenkamp: Lehrbuch der Pädagogischen Diagnostik, 3. unveränderte Aufl., Weinheim: Beltz 1995
- ¹⁵ E. W. Kleber: Diagnostik in pädagogischen Handlungsfeldern, Wien: Juventa 1992
- ¹⁶ vgl. K. Ingenkamp: Die Test-Aversion des deutschen Intellektuellen, in: K. Ingenkamp (Hrsg.): Diagnostik in der Schule: Beiträge zu Schlüsselfragen der Schülerbeurteilung, Weinheim: Beltz 1989, S. 153 - 179.

Literatur

- Diehl, J.M., H.U. Kohr: Deskriptive Statistik, 11. Aufl., Eschborn: Klotz 1994
- Hopkins, K.D., J.C. Stanley, R.B. Hopkins: Educational and Psychological Measurement and Evaluation, seventh edition, Englewood Cliffs, New Jersey: Prentice Hall 1990
- Ingenkamp, K.: Die Test-Aversion des deutschen Intellektuellen, in: K. Ingenkamp (Hrsg.): Diagnostik in der Schule: Beiträge zu Schlüsselfragen der Schülerbeurteilung, Weinheim: Beltz 1989, S. 153–179
- Ingenkamp, K.: Lehrbuch der Pädagogischen Diagnostik, 3. unveränderte Aufl., Weinheim: Beltz 1995
- Klauer, K.J.: Neuere Entwicklungen in der pädagogisch-psychologischen Diagnostik, in: Diagnostica 40 (1994) 3, S. 244–261
- Kleber, E.W.: Diagnostik in pädagogischen Handlungsfeldern, Wien: Juventa 1992
- Kubinger, K.D.: Einführung in die psychologische Diagnostik, Weinheim: Psychologie-Verlags-Union 1995
- Plaum, E.: Einführung in die Psychodiagnostik, Darmstadt: Primus-Verlag 1996
- Testzentrale des Berufsverbandes Deutscher Psychologen (Hrsg.): Testkatalog 1996/97, Göttingen: Hogrefe 1996

Anmerkungen

- ¹ nach B. Hilberg, Schulpraktikumsbericht, Justus-Liebig-Universität Gießen, Institut für Arbeits-, Berufs- und Wirtschaftspädagogik, Gießen 1976, S. 14–16.
- ² vgl. J. C. Stanley, K. D. Hopkins, Educational and Psychological Measurement and Evaluation, Englewood Cliffs, New Jersey 1972, S. 23.
- ³ vgl. G. V. Glass, J. C. Stanley, Statistical Methods in Education and Psychology, Englewood Cliffs, New Jersey 1970, S. 59 f.
- ⁴ Zum Vergleich: Im Lesereifetest, Tabelle 4, S. 18, beträgt die Interquartilweite 24,5 Rohwerte: Die mittleren 50% der Rohwerte verteilen sich auf 24,5 Rohwerte. Q_1 beträgt 69,5 und Q_3 beträgt 94.
- ⁵ nach J. P. Guilford, Fundamental Statistics in Psychology and Education, 2. ed., New York 1950, S. 96 f.
- ⁶ Wendepunkte sind die Punkte, bei denen die Krümmung einer Kurve sich ändert. – Durchfährt ein Motorradfahrer eine S-Kurve, ist der Wendepunkt dort, wo der Fahrer sein Gewicht zum Durchfahren der Kurve von der einen zur anderen Seite verlagert.
- ⁷ vgl. im folgenden (S. 48–52) E. P. Lovejoy, Statistics for math haters, New York 1975, S. 141–146.
- ⁸ Hier liegt die Annahme der klassischen Testtheorie zugrunde, daß die Unzulänglichkeit des Meßinstrumentes die anormale Verteilung verursacht. Bei der nicht-klassischen Testtheorie wird diese Annahme in Frage gestellt: Die anormale Verteilung kann auch das wahre Leistungsbild widerspiegeln.
- ⁹ vgl. R. L. Ebel, Essentials of Educational Measurement, Englewood Cliffs, New Jersey 1972, S. 294.
- ¹⁰ vgl. G. Rapp, Messung und Evaluierung von Lernergebnissen in der Schule, Bad Heilbrunn/Obb. 1974, S. 164.
- ¹¹ R. Weiß, Grundintelligenztest CFT 3-Skala 3, Braunschweig 1971.
- ¹² W. Horn, Prüfungssystem für Schul- und Bildungsberatung P-S-B, Göttingen 1969.
- ¹³ G. A. Lienert, Die Drahtbiegeprobe als standardisierter Test, 2. verb. Aufl., Göttingen 1967.
- ¹⁴ vgl. G. Clauß, H. Ebner, Grundlagen der Statistik für Psychologen, Pädagogen und Soziologen, ver. u. erw. Aufl., Berlin 1974, S. 116 ff.
- ¹⁵ aus Englisch valid und validity (gültig, zwingend bzw. Gültigkeit, Stichhaltigkeit) in das Deutsche zu valide und Validität übertragen.
- ¹⁶ vgl. R. Weiß, Grundintelligenztest CFT 3-Skala 3: Handanweisung für die Durchführung, Auswertung und Interpretation, Braunschweig 1971, S. 9.
- ¹⁷ vgl. R. Weiß, Grundintelligenztest CFT 3-Skala 3: Handanweisung für die Durchführung, Auswertung und Interpretation, Braunschweig 1971, S. 9 und 35 f.
- ¹⁸ vgl. J. C. Stanley, K. D. Hopkins, Educational and Psychological Measurement and Evaluation, Englewood Cliffs, New Jersey 1972, S. 111 f.
- ¹⁹ „Mongolismus“: eine angeborene geistige Behinderung.
- ²⁰ aus Englisch reliable und reliability (zuverlässig, glaubwürdig bzw. Zuverlässigkeit, Verlässlichkeit) in das Deutsche zu reliabel und Reliabilität übertragen.

- ²¹ Zur Erklärung einer „notwendigen“ und „hinreichenden“ Bedingung: Wenn eine Straße naß ist, dann ist dies eine notwendige Bedingung dafür, daß es geregnet hat. Es ist aber keine hinreichende Bedingung, denn es könnte auch jemand die Straße mit einem Wasserschlauch naß gespritzt haben.
- ²² Beim Münzenwurf errechnet sich die Standardabweichung nach der Binominalverteilung: $s = \sqrt{N \cdot p \cdot q}$ (vgl. G. Clauß, H. Ebner, Grundlagen der Statistik für Psychologen, Pädagogen und Soziologen, ver. u. erw. Aufl., Berlin 1974, S. 144). N ist die Zahl der Ziehungen oder Würfe pro Reihe. Hier ist $N = 100$. p ist die Wahrscheinlichkeit des Auftretens des einen Zustandes, q die des anderen Zustandes; hier ist $p_{\text{wappen}} = 0,5$ und $q_{\text{zahl}} = 0,5$. Somit ist $s = \sqrt{100 \cdot 0,5 \cdot 0,5} = 5$. Gemäß der Normalverteilungskurve wird in 68 der 100 Versuchsreihen zu je 100 Münzwürfen die Rohwertzahl zwischen 45 und 55 liegen (Bereich von $\bar{x} - 1 \cdot s$ bis $\bar{x} + 1 \cdot s$).
- ²³ vgl. G. A. Lienert, Testaufbau und Testanalyse, 3. erg. Aufl., Weinheim 1969, S. 221.
- ²⁴ vgl. J. C. Stanley, K. D. Hopkins, Educational and Psychological Measurement and Evaluation, Englewood Cliffs, New Jersey 1972, S. 125.
- ²⁵ vgl. R. Weiß, Grundintelligenztest CFT 3-Skala 3: Handanweisung für die Durchführung, Auswertung und Interpretation, Braunschweig 1971, S. 8.
- ²⁶ vgl. G. A. Lienert, D-B-P Die Drahtbiegeprobe als standardisierter Test: Handanweisung für die Durchführung und Auswertung, 2. verb. Aufl., Göttingen 1967, S. 17.
- ²⁷ vgl. G. A. Lienert, Testaufbau und Testanalyse, 3. erg. Aufl., Weinheim 1969, S. 227.
- ²⁸ vgl. G. A. Lienert, Testaufbau und Testanalyse, 3. erg. Aufl., Weinheim 1969, S. 227.
- ²⁹ vgl. R. L. Ebel, Essentials of Educational Measurement, Englewood Cliffs, New Jersey 1972, S. 415 f.
- ³⁰ vgl. G. A. Lienert, D-B-P Die Drahtbiegeprobe als standardisierter Test: Handanweisung für die Durchführung und Auswertung, 2. verb. Aufl., Göttingen 1967, S. 17.
- ³¹ vgl. R. Weiß, Grundintelligenztest Skala 3 – CFT 3: Handanweisung für die Durchführung, Auswertung und Interpretation, Braunschweig 1977, S. 8.
- ³² vgl. G. A. Lienert, D-B-P Die Drahtbiegeprobe als standardisierter Test: Handanweisung für die Durchführung und Auswertung, 2. verb. Aufl., Göttingen 1967, S. 17.
- ³³ vgl. R. Weiß, Grundintelligenztest Skala 3 – CFT 3: Handanweisung für die Durchführung, Auswertung und Interpretation, Braunschweig 1971, S. 22 und 45, nach der Gesamtnormentabelle.
- ³⁴ vgl. G. A. Lienert, D-B-P Die Drahtbiegeprobe als standardisierter Test: Handanweisung für die Durchführung und Auswertung, 2. verb. Aufl., Göttingen 1967, S. 17.
- ³⁵ vgl. G. A. Lienert, Testaufbau und Testanalyse, 3. erg. Aufl., Weinheim 1969, S. 455.
- ³⁶ vgl. K. Heller (Hrsg.), Leistungsbeurteilung in der Schule, 2. durchgeseh. Aufl., Heidelberg 1975, S. 25 f. und S. 137 ff.

- ³⁷ vgl. G. Clauß, H. Ebner, Grundlagen der Statistik für Psychologen, Pädagogen und Soziologen, ver. u. erw. Aufl., Berlin 1974, S. 298 ff.
- ³⁸ vgl. J. C. Stanley, K. D. Hopkins, Educational and Psychological Measurement and Evaluation, Englewood Cliffs, New Jersey 1972, S. 268 f.
- ³⁹ nach B. Hilberg, Schulpraktikumsbericht, Justus Liebig-Universität Gießen, Institut für Arbeits-, Berufs- und Wirtschaftspädagogik, Gießen 1976, S. 37–39.
- ⁴⁰ vgl. hierzu auch Kap. 11.4 Distraktorenanalyse.
- ⁴¹ nach Industrie- und Handelskammer Mittlerer Neckar (Hrsg.), PAL-Aufgabenbank, Elektroberufe (Energietechnik) Fachkunde Teil 2, Konstanz 1975, S. 13.
- ⁴² Diese Aufgabe ist dem Kap. 11.4 Distraktorenanalyse entnommen.
- ⁴³ B. S. Bloom et al. (Hrsg.), Taxionomie von Lernzielen im kognitiven Bereich, Weinheim 1972, S. 149.
- ⁴⁴ zu weiteren Beispielen s. neben Bloom 1972, G. P. Bunk, Revision der kaufmännischen Berufsausbildung, Heidelberg 1974, S. 140 ff.
- ⁴⁵ vgl. G. A. Lienert, Testaufbau und Testanalyse, 3. erg. Aufl., Weinheim 1969, S. 331.
- ⁴⁶ vgl. R. L. Thorndike, E. Hagen, Measurement and Evaluation in Psychology and Education, 3. ed., New York 1969, S. 581.
- ⁴⁷ vgl. J. Wendeler, Standardarbeiten, 5. überarbeitete Aufl., Weinheim 1973, S. 69.
- ⁴⁸ vgl. E. Hoffmann, Zur Frage von Punktbewertungsrichtlinien, in: Berufsbildung in Wissenschaft und Praxis 6 (1977) 2, S. 24.

Literaturverzeichnis*

- Birkel, P., Glossar wichtiger testtheoretischer Begriffe, in: Test Informationen – Hinweise zur Auswahl von Tests, Beltz Test Gesellschaft, Weinheim 1976, S. 27–51.
- Bloom, B. S., et al. (Hrsg.), Taxonomie von Lernzielen im kognitiven Bereich, Weinheim 1972.
- Bunk, G. P., Revision der kaufmännischen Berufsausbildung, Heidelberg 1974.
- Clauß, G., Ebner, H., Grundlagen der Statistik für Psychologen, Pädagogen und Soziologen, ver. u. erw. Aufl., Berlin 1974.
- Diehl, J. M., Kohr, H. U., Statistische Methoden: Band I Deskriptive Statistik, 3. verb. Aufl., Fachbereich Psychologie, Justus Liebig-Universität Gießen o. J.
- Ebel, R. L., Essentials of Educational Measurement, Englewood Cliffs, New Jersey 1972.
- Glass, G. V., Stanley, J. C., Statistical Methods in Education and Psychology, Englewood Cliffs, New Jersey 1970.
- Guilford, J. P., Fundamental Statistics in Psychology and Education, 2. ed., New York 1950.
- Heller, K. (Hrsg.), Leistungsbeurteilung in der Schule, 2. durchgeseh. Aufl., Heidelberg 1975.
- Hilberg, B., Schulpraktikumsbericht, Justus Liebig-Universität Gießen, Institut für Arbeits-, Berufs- und Wirtschaftspädagogik, Gießen 1976.
- Hoffmann, E., Zur Frage von Punktbewertungsrichtlinien, in: Berufsbildung in Wissenschaft und Praxis 6 (1977) 2, S. 23–25.
- Horn, W., Prüfungssystem für Schul- und Bildungsberatung P-S-B, Göttingen 1969.
- Industrie- und Handelskammer Mittlerer Neckar (Hrsg.), PAL-Aufgabenbank: Elektroberufe (Energietechnik) Fachkunde Teil 2, Konstanz 1975.
- Ingenkamp, K. (Hrsg.), Tests in der Schulpraxis, Weinheim 1971.
- Lienert, G. A., D-B-P Die Drahtbiegeprobe als standardisierter Test, 2. verb. Aufl., Göttingen 1967.
- Lienert, G. A., Testaufbau und Testanalyse, 3. erg. Aufl., Weinheim 1969.
- Lovejoy, E. P., Statistics for math haters, New York 1975.
- Rapp, G., Messung und Evaluierung von Lernergebnissen in der Schule, Bad Heilbrunn/Obb. 1974.
- Stanley, J. C., Hopkins, K. D., Educational and Psychological Measurement and Evaluation, Englewood Cliffs, New Jersey 1972.
- Thorndike, R. L., Hagen, E., Measurement and Evaluation in Psychology and Education, 3. ed., New York 1969.
- Weiß, R., Grundintelligenztest CFT 3 -Skala 3, Braunschweig 1971.
- Wendeler, J., Standardarbeiten, 5. überarbeitete Aufl., Weinheim 1969.

* Siehe auch das Verzeichnis der Literatur auf S. 181

Anhang

Lösungen zum Test Vitamine und Mineralstoffe, S. 10f.

- Aufgabe 1: b)
Aufgabe 2: b); c)
Aufgabe 3: fettlösliche Vitamine; wasserlösliche Vitamine
Aufgabe 4: Mengenelemente; Spurenelemente
Aufgabe 5: Obst; Gemüse
Aufgabe 6: c)
Aufgabe 7: Mangelerscheinungen bzw. Mangelkrankungen (Avitaminosen) verursacht werden.
Aufgabe 8: a); c); d); f)
Aufgabe 9: c) fettlösliche Vitamine lösen sich nicht in Wasser auf.
Aufgabe 10: Ascorbinsäure
Aufgabe 11: µg bis mg, auch die Angabe, daß der Bedarf tausend bis millionenfach kleiner ist als der von Eiweiß, Fett u. KH
Aufgabe 12: Der Körper kann diese Stoffe nicht selbst aufbauen, sie müssen mit der Nahrung zugeführt werden.
Aufgabe 13: z.B.: Ca, Mg, K, Na, Cl, S, P, Fe
Aufgabe 14: z.B.: Cu, Mn, Zn, Co, Mo, Si, Se, J, F
Aufgabe 15: Milch
Aufgabe 16: c)–1); b)–2)
Aufgabe 17: Ja, in Wasser
Aufgabe 18: Kinder haben erhöhten Bedarf, da die Mineralstoffe mit Baustoffe des Körpers sind.
Aufgabe 19: Vitamine und Mineralstoffe befinden sich vor allem in der Schale
Aufgabe 20: Provitamin

Sachregister

Abweichungs-IQ s. IQ-Skala

Abweichungsquadrate 39f.

aktueller Rohwert 103f.

aktuelle Varianz 106ff.

Antworten

–, alternative s. Richtig-Falsch Aufgaben bzw. Mehrfachwahlaufgaben

–, freie 144

–, gebundene 144

arithmetischer Mittelwert 25 ff., 172

Aufgabenanalyse 128, 132 ff., 138 ff.

Aufgabenarten 144

Aufgabenschwierigkeit s. Schwierigkeit

Aufgabenselektion 134 ff., 141

Aufgabenstamm 146

Aufgabentrennschärfe s. Trennschärfe

Auswahlantwortaufgaben 144

Auswertungsobjektivität 125f.

Axiom 122

Benotung s. Notengebung

beobachtete Varianz s. aktuelle Varianz

Centil-Skala 59, 61 f., 64 f., 77, 175

concurrent validity s. Übereinstimmungsvalidität

confidence limits 119

content validity s. Inhaltsvalidität

cumulierte Häufigkeit 21 f., 169

Diskriminationsindex D 132f.

Distraktorenanalyse 136ff.

Durchführungsobjektivität 124f.

Eichung 73 f., 81

Ergänzungsaufgaben 152ff.

Essay-Test s. Kurzantwortaufgaben (Kurzaufsatzaufgaben)

Expertenurteil 98

Falschantworten s. Distraktorenanalyse

Fehler s. Meßfehler, Standardmeßfehler

Fehlervarianz 106ff.

Flanagan-Formel 111, 117

Flächentransformation 66f.

Genauigkeit s. Reliabilität

grafische Normen 79 ff.

Grundausswertung 16 ff., 165 ff.

Gültigkeit s. Validität

Gütekriterien s. Objektivität, Reliabilität u. Validität

Halbierungsreliabilität s. Testhalbierungsreliabilität

Häufigkeit

–, cumulierte 21 f., 169

Häufigkeitsverteilung 17 ff., 163 ff.

–, Primäre – 17 ff., 163 ff.

–, Sekundäre – 17

Informeller Test 9, 81, 98

Inhaltsvalidität 98f.

Intelligenzquotient s. IQ-Skala

Interpretationsobjektivität 124, 126f.

Interquartilweite 34

IQ-Begriff, älterer 60f.

IQ-Skala 59, 61 f., 64 f., 175

Itemanalyse s. Aufgabenanalyse

Itemschwierigkeit s. Schwierigkeit

Itemtrennschärfe s. Trennschärfe

Klassische Testtheorie 122 ff., 179 f.

Konsistenzanalyse 112 ff., 117

Konstrukt 100

Konstruktvalidität 100f.

Korrelation

–, Interpretation 91 ff.

–, Konzept 83 ff.

– und Kausalität 93f.

Korrelationskoeffizient

–, Pearson-Produkt-Moment 86, 176

–, punktbiserialer 131

kriterienbezogene Validität 99f.
 kritische Differenz 121f.
 Kuder-Richardson-Formel 20 u.
 21 112, 113, 117
 Kurzantwortaufgaben (Kurzaufsatz-
 aufgaben) 155 ff.

 Lehrplanvalidität 98
 Leistungsmessung
 –, kriterienbezogene -123
 –, normbezogene 123
 Lernziele 98
 lineare Transformation 67, 68 f., 70

 Maßkorrelationskoeffizient 86, 176
 Median 20 ff., 168 ff.
 Messung 104 ff.
 Meßfehler 104 ff.
 Mehrfachwahlaufgabe 136 ff., 146 ff.
 –, extrinsische, intrinsische Mehrdeu-
 tigkeit 136 ff.
 Mittelwert
 –, arithmetischer 25 ff., 172
 Modalwert 24 ff., 170 f.
 multiple-choice Aufgabe s. Mehr-
 fachwahlaufgabe

 Normalverteilung 28 f., 56
 Normen 73
 –, tabellarische 73 ff.
 –, grafische 79 ff.
 Normierung 81 f.
 Noten
 –, nach der Normalverteilung: Pro-
 blematik 160 ff.
 –, Notengebung 157
 –, Schulnoten: SN-Skala 158
 –, Stasix-Skala 158
 –, Verteilungsschemata 157 ff.
 –, 100-Punkte-Schlüssel der In-
 dustrie- und Handelskam-
 mern 160

 Objektivität
 –, Auswertungs- 125 f.
 –, Durchführungs- 125

–, Interpretations- 126 f.
 –, Testobjektivität: Begriff 124 f.
 Options s. Auswahlantwortaufgaben
 Ordnungsantwortaufgaben 144
 Paralleltestreliabilität 115 f., 117
 Pearson-Produkt-Moment Korrela-
 tionskoeffizient 86, 176
 Perzentil 33, 43 ff., 64, 173 f.
 prädikative Validität (predictive vali-
 dity) s. Vorhersagevalidität
 prognostische Validität s. Vorhersa-
 gevalidität
 Prozentrang 45 ff., 54, 64 f., 65 f., 173 f.
 Prozentrangnorm 65 f.
 punktbiseriale Korrelation 131

 Quartil 33, 172
 – abweichung 33 ff., 41, 172

 Range 32
 Rangfolge 14 ff., 18, 164 f.
 Ratewahrscheinlichkeit bei Aufgabe-
 antworten 146, 148, 149
 Reliabilität
 –, Begriff 102 f., 106 f., 125
 –, Konsistenzanalyse 112 ff., 117
 –, Paralleltest 115 f., 117
 –, Retest 115 f., 117
 – skoeffizient 106 f., 108 ff., 116 f.
 –, Testhalbierung 110 ff., 117
 Reliabilitätsverbesserung 117 ff.
 Retest-Reliabilität 115 f., 117
 Richtig-Falsch-Aufgaben 144 ff.
 Rohwert 14
 –, aktueller 103 f.
 –, wahrer 103 f.

 schiefe Verteilung 28 f.
 Schulnoten 157 ff.
 Schulnotenskala 157 f.
 Schwierigkeit 118, 134 f.
 Schwierigkeitsgrad 128 f.
 Semi-Interquartilsrange 34
 Skala
 –, IQ 59, 61 f., 64 f., 175 f.
 –, C 59, 61 f., 64 f., 175 f.
 –, T 59 f., 64 f., 175 f.

- , Z 59, 64 f., 79 f., 175 f.
- Spearman-Brown-Formel 110 f., 117
- split-half-Reliabilität s. Testhalbierungsreliabilität
- Stamm (der Aufgabe) 146
- Standardabweichung 34 ff., 173
- Standardisierung 73 f., 81 f.
- Standardmeßfehler 105 ff.
- , Vertrauensintervall 118 ff.
- Standardnormalverteilung 48 ff., 53 f., 63 f.
- Standardwerte z 55 ff., 63 f., 174 f.
- Standardwertskalen 58 ff., 61 f., 175
- Stanine 67 ff.
- Stasix 70 f.
- , Notenskala 158
- Sten 71
- symmetrische Verteilung 28 f.
- tabellarische Normen 73 ff.
- Test
 - , Eichung, Normierung, Standardisierung 73 f., 81 f.
 - halbierungsreliabilität 110 ff., 117
 - informeller 9, 81, 98
 - objektivität s. Objektivität
 - reliabilität s. Reliabilität
 - validität s. Validität
- Testaufgaben 144 ff.
- , Ergänzungsaufgaben 152 ff.
- , Kurzantwortaufgaben (Kurzaufsatzaufgaben) 155 ff.
- , Mehrfachwahlaufgaben 146 ff.
- , Richtig-Falsch-Aufgaben 144 ff.
- , Umordnungsaufgaben 151 f.
- , Zuordnungsaufgaben 149 ff.
- Testtheorie
 - , klassische 122 ff., 179 f.
 - , Axiome der klassischen – 122 f.
- Transformation 54 ff., 58 ff., 65 f.
- , Flächen- 66 f., 69
- , lineare- 58, 67 f., 71
- Trennschärfe 117, 129 ff., 134 f.
- koeffizient 129 ff.
- T-Skala 59, 64 f., 175 f.

- Übereinstimmungsvalidität 99
- Umordnungsaufgaben 151 f.
- unimodale Verteilung 28 f.
- Urliste 14, 163
- Validität
 - , Begriff 98, 102 f., 124 f.
 - , Inhalts- 98 f.
 - , Konstrukt- 100 f.
 - , kriterienbezogene 99 f.
 - , Lehrplan- 98
 - , Übereinstimmungs- 99
 - , Vorhersage- 100
- Variabilität 32 ff., 172 f.
- smaße 32 ff., 172 f.
- Varianz 39 ff., 173
- , aktuelle 106 ff.
- , Fehler- 106 ff.
- , wahre- 106 ff.
- Variationsweite 32 f., 172
- Verteilung
 - , normal 28, 47 ff., 56
 - , schief, symmetrisch 28 f.
 - , Standardnormal- 48 ff., 53 f., 64
 - , unimodal 28 f.
 - z- 48 ff.
- Vertrauensintervall
 - , Standardmeßfehler 118 ff.
- Vorhersagevalidität 100
- wahrer Rohwert 103 ff.
- wahre Varianz 106 ff.
- Zensierung s. Notengebung
- Zensur s. Noten
- zentrale Tendenz 20 ff., 29 ff.
- Z-Skala 59, 64 f., 80, 175 f.
- z-Standardwerte 55 ff., 174 f.
- z-Tabelle 50
- z-Transformation 55, 174
- z-Verteilung 48 f.
- Zuordnungsaufgaben 149 ff.
- Zuverlässigkeit s. Reliabilität

Aus Rezensionen der 1. Auflage:

„In wirklich verständlicher Sprache werden an Hand leicht nachrechenbarer Beispiele alle für die Konstruktion standardisierter Tests wesentlichen Begriffe und Vorgangsweisen anschaulich dargestellt. Ein sehr brauchbarer Behelf für Studierende der Psychologie und Pädagogik und für Lehrer!“

(Erziehung und Unterricht)

„Abschließend glaube ich sagen zu dürfen, daß Pädagogen, die daran interessiert sind, ihre Leistungserfassung objektiver zu gestalten und gängige Leistungstests angemessen zu beurteilen und einsetzen zu können, in diesem didaktisch geschickt aufgebauten Buch einen hilfreichen, praxisnahen Leitfaden finden.“

(Erziehungstherapie)



Franz Steiner Verlag Stuttgart