Schurig, M. & Gebhardt, M. (2022). Theoretische Grundlagen von Messungen und Tests. In M. Gebhardt, D. Scheer & M. Schurig (Hrsg.), *Handbuch der sonderpädagogischen Diagnostik. Grundlagen und Konzepte der Statusdiagnostik, Prozessdiagnostik und Förderplanung* (S. 233-246). Regensburg: Universitätsbibliothek. https://doi.org/10.5283/epub.53149

Theoretische Grundlagen von Messungen und Tests

Michael Schurig & Markus Gebhardt

1 Messen und Testen

»We see far too many good ideas slaughtered by poor measurement practices […]«

(Reynolds, 2010)

Messen und Testen sind essentielle Bestandteile empirischer Forschung. **Messen** beschreibt, wie einem beobachteten Phänomen ein meist numerischer Wert gegeben wird. Das Messen selbst ist dabei mit theoretischen Annahmen zu Konstrukten, z. B. dem Verhalten, und deren Messbarkeit verbunden. Ziel ist es ein Konstrukt zu operationalisieren und beispielsweise »mehr« oder »weniger« störendes Verhalten zu messen. Ein anderes Beispiel ist die Erfassung der kognitiven Fähigkeiten eines Kindes, um die Ausprägung der fluiden Intelligenz im Vergleich zu einer Norm abzubilden. Ebenso kann man auch Gruppenunterschiede messen, um zu zeigen, dass die Interventionsgruppe doppelt so starke Leistungen im Leseverstehen erreicht wie die Kontrollgruppe oder auch um Benachteiligung im Bildungssystem zu erfassen, um über eine von Armut bedrohte Gruppe zu sprechen.

Der Begriff **»Test«** ist wissenschaftlich weiter gefasst zu verstehen, als es in der Alltagssprache erscheint. Nach Lienert und Raatz (1998, S. 7) umfassen Tests unter anderem Verfahren zur Untersuchung von Persönlichkeitsmerkmalen, den Vorgang der Durchführung der Untersuchung und mathematisch-statistische Prüfverfahren. Somit werden unter dem Begriff Test sowohl die möglichen Zusammenfassung(-en) oder Aggregation(-en) von einzelnen Beobachtungen oder Testitems (Aufgaben) zu einem sogenannten Score oder Messwert als auch die Weiterverarbeitung dieser Scores in statistischen Modellen erfasst. In diesem Beitrag steht die Erstellung von Scores gebildet aus einer Gruppe standardisierter Items im Mittelpunkt (vgl. Rost, 2004).

Den Rahmen für beide Ansätze in quantitativer Tradition geben die Messtheorie (z. B. Orth, 1974) und Testtheorien (z. B. Fischer, 1974; Lienert & Raatz, 1998; Rost, 2004) vor. Sowohl die Mess- als auch die Testtheorie beschäftigen sich mit statistischen Messmodellen. Die Unterschiede liegen dabei im empirischen Ausgangspunkt (Steyer & Eid, 2001; S. V).

2 Beispiele zur Anwendung der Mess- und Testtheorie:

Zur Messtheorie: »Ich fühle mich in meiner Klasse wohl.« ist ein häufiges Item in einem Fragebogen zur Messung des latenten Konstrukts Klassenklima. Der Begriff der Latenz meint dabei vor allem, dass das Konstrukt theoretisch vorhanden, aber nicht direkt beobachtbar ist. Häufige Antwortkategorien lauten bei einer vierstufigen Skala »stimme gar nicht zu«, »stimme eher nicht zu«, »stimme zu« und »stimme vollkommen zu«. Diese Werte werden für die Weiterverarbeitung in Kategorien von 1 bis 4 operationalisiert. Die Skala kann aber ebenso mit einer Wertausprägung »teils/teils« ergänzt werden. In diesem Fall können die Kategorien Ausprägungen von 1-5 annehmen. Je nach Fragestellung kann eine vierstufige oder fünfstufige Antwortskala angemessener sein. Wenn eine Normalverteilung erreicht werden soll, was oft wünschenswert ist, so sind Werte von 1-5 angemessener, da die Ergebnisse hierbei leichter eine Glockenkurve approximieren. Wenn man jedoch kontrastieren will und es auf Rangzuweisungen zwischen Schüler:innen ankommt, ist die vierstufige Skala angemessener. Hierbei zwingt man die Befragten eine positive oder negative Entscheidung zu treffen. Bei der Festlegung des Niveaus der Wertskala (Skalenniveau) adressiert man bereits die mit diesem Niveau

- möglichen Transformationen, also Umrechnungen zur besseren Interpretierbarkeit, z. B. lineare Transformation,
- die verwendbaren Lagemaße, welche die Messwerte einer Verteilung zusammenfassend repräsentieren, z. B. der Mittelwert oder Median,
- die zu verwendenden Verteilungsmaße, also die Streuungen, wie z. B. Häufigkeiten oder die Standardabweichung und
- die ableitbaren Verhältnisse (z. B. »gleich« oder »ungleich«, »mehr« oder »weniger«).

Zur Testtheorie: Wenn auf die Fähigkeit geschlossen werden soll Additionsaufgaben im Zahlenbereich bis 20 zu lösen, könnten zehn Additionsaufgaben verwendet werden. Die Antworten könnten als »richtig« (1) oder »falsch« (0) codiert werden. Alternativ könnten, wenn auch korrekte Lösungswege beobachtet werden, die Codierungen »Antwort richtig und Lösungsweg richtig« (2) und »Lösungsweg richtig aber Antwort falsch« (1) und »Lösung richtig, aber Antwortweg falsch oder beides falsch« (0) codiert werden.

Wie werden die einzelnen Werte aber zusammengefasst? Darf man die Werte aufsummieren oder ist es eventuell angemessener einen Mittelwert zu bilden? Wie wird dabei mit fehlenden Werten umgangen? Hierbei kann man entweder nicht bearbeitete Items bei einem Test als falsch bewerten, da man annimmt die Testperson hat das Item absichtlich übersprungen und kann es nicht lösen. Oder man berücksichtigt diese Aufgaben bei der Bildung des Scores nicht, da man an sich keine Information über das Item hat. In der Summe sind durch fehlende Werte die maximal erreichbaren Punkte im Einzelfall zwischen den Testteilnehmern unterschiedlich (z. B. gilt bei einer Codierung von 0/1 und 10 Items ein Maximum von 10, wenn aber ein Item nicht beantwortet wird, dann ist das potenzielle Maximum 9). Um dies zu lösen, kann der Anteil der gelösten Items als Score verwendet werden. z. B. ergeben 7 von 10 gelösten Aufgaben einen Summenscore von 7, respektive einen Mittelwert von 0.7. Wenn eine Aufgabe nicht beantwortet wird, so ist der Summenscore bei 6 von 9 gelösten Aufgaben gleich 6 und der Anteil gelöster Aufgaben ist folglich 0.667. Der Mittelwert sollte in diesem Fall korrigiert werden, indem die Anzahl der verfügbaren Werte im Nenner berücksichtigt werden anstatt die Anzahl der potentiell verfügbaren Werte (n_i) zu wählen.

$$Mittelwert = (\frac{Wert_1 + Wert_{...} + Wert_{n_i}}{n_i \, oder \, Anzahl \, der \, verfuegbaren \, Werte}) \quad (1)$$

Insbesondere für Erhebungen auf der Basis von auszufüllenden Fragebögen besteht häufiger das Problem von nicht bearbeiteten Items. Daher besteht immer die Frage, wie viele fehlende Werte vorliegen dürfen, damit überhaupt ein Mittelwert gebildet werden darf. Einige Quellen geben an, dass mindestens zwei Drittel der Items vorliegen sollten, damit ein Mittelwert gebildet werden kann (vgl. Lord, 1974). Eine weitere Alternative für Forschungsarbeiten ist das statistische Auffüllen fehlender Werte aus den beobachtbaren Zusammenhängen (Lüdtke et al., 2007).

Bei der Erstellung eines Scores kann es problematisch sein, wenn Items unterschiedliche statistische Relevanz (Korrelation des Items mit der Skala; r_{it}; Trennschärfe) oder Schwierigkeiten (Wahrscheinlichkeit ein Item zu lösen; p_i) aufweisen. Gehen beispielsweise auch besonders schwere Items in die Berechnung eines Wertes ein, obwohl die Lösung von schwachen Lernenden nie erreicht wird? Dies bedeutet, dass die Anzahl der potenziell nützlichen Items für schwache Lernende besonders gering ist und die Testlänge insgesamt für Lernende mit schwachen Leistungen kürzer ist, da diese Items keine diagnostischen Informationen liefern (vgl. Renner, in diesem Band). Je nach Schwierigkeiten der Items ist eine solche Skala für Kinder mit schwachen Leistungen ungenauer. Mithilfe der Testtheorie kann man diese Fragen beantworten.

Man kann sagen, dass beide Theorien die Frage adressieren, wie etwas von dem ich annehme, dass es existiert, aber für das es kein abzählbares Maß gibt, formal beschrieben werden kann. Denken wir uns einen dreidimensionalen Quader. Die Messtheorie begründet die Beschriftung unseres Maßbandes mit dem wir die Seitenlängen und ggf. die Querschnitte der Flächen bestimmen. Durch die Messtheorie werden den direkt beobachtbaren Eigenschaft des Quaders (also z. B. den Seitenlängen) numerische Werte gegeben. Zum Beispiel hat das Klassenklima hat kein natürliches Maß. Es braucht die Messtheorie zur Festlegung der Metrik der beobachtbaren Eigenschaften, also der Items in z. B. in Werten von 1-5. Die Testtheorie begründet, welche Werte wir verwenden sollten um Eigenschaften des Quaders formal zu beschreiben (z. B. die Formel des Volumens = Seitenlänge a mal Seitenlänge b mal Seitenlänge c). Für die Bewertung das Klassenklima auf Individualebene könnte man also z. B. einen Mittelwert aus allen Items bilden.

2.1 (Repräsentative) Messtheorie

Messmodelle beschreiben die Beziehung zwischen Annahme und Beobachtung – also Theorie und Empirie – und geben dem untersuchten Gegenstand eine formale, mathematische, Struktur. Die Angabe »stimmt nicht« wird eine 0 und eine richtige Lösung wird eine 1. Diese numerischen Werte sind dabei oft notwendige Voraussetzungen der statistischen Prüfung eines angenommenen, also theoretisierten, Modells (Steyer & Eid, 2001). Die Messtheorie ist also Bedingung einer Übertragung von der Theorie in die quantitative empirische Praxis. Es existieren unterschiedliche Messtheorien, die klassische und die operationale und die repräsentative (Orth, 1974). Erstere geht von einer perfekten Repräsentierung aus (z. B. Anzahl Fenster in einem Raum), die zweite meint die Bestimmung des Messgegenstands durch die Messung. Dies kann z. B. die operationale Definition der Armut umfassen. Nach einer üblichen Definition ist man von Armut bedroht, wenn das monatliche Äquivalenzeinkommen weniger als 50 % des

Tabelle 1: Skalenniveaus (Schurig, 2017; vgl. Stevens, 1946)

Skalenniveau	Definition	Mögliche Lagemaße	Mögliche Verteilungs-funktionen	Verwendbare Statistiken	Beispiele
Nominalskala	Reine Betitelung, keine Ordnung möglich	Modus	Häufigkeiten	Gleichheit/ Ungleichheit	Geschlecht, Familienstand
Ordinalskala	Ordnung, aber keine Interpretation der Abstände möglich	Median	Perzentile/ Spannweite	Größer/ Kleiner	Schulnoten, Straßennummern
Intervallskala	Wie Ordinalskala, aber Interpretation der Abstände möglich	Arithmetisches Mittel	Varianz/ Standardabweichung	Unterschied/ Distanz	Zeiträume
Ratio- oder Verhältnisskala	Wie Intervallskala, aber Sinnvoller Nullpunkt vorhanden	Geometrisches Mittel und harmonisches Mittel	Variablilität	Verhältnis	Blutdruck, Länge, Gewicht, Alter

Medians des Landeseinkommens beträgt (OECD, 2022). Eine perfekte Repräsentation von Armut durch direkte Messungen wie beispielsweise durch die Zahl von Haushaltsgeräten oder anderen Dingen des alltäglichen Lebens (z. B. zeigt der Besitz eines Rasenmähers an, dass ein Rasen vorliegt, also meist ein Haus bewohnt wird) wäre sehr viel aufwendiger und nicht genauer.

Die **repräsentative Messtheorie** meint die möglichst strukturerhaltende Abbildung von Beobachtungen in numerische Werte und ist der relevanteste Ansatz in den Geistes- und Sozialwissenschaften. Begründet wurde die moderne repräsentative Messtheorie durch den Psychologen Stevens (1946, 1957). Er etablierte hierzu eine Hierarchie von Skalenniveaus, die definiert, welche mathematischen Operationen, welche Transformationen und welche Interpretationen zulässig sind. Die wichtigsten Skalentypen sind die Nominal-, die Ordinal-, die Intervallskala. In den meisten Anwendungen werden Intervall- und darüberhinausgehende Skalenniveaus die hier nicht vorgestellt werden vergleichbar behandelt. Die angemessenen Verteilungsfunktionen sowie zulässigen Verhältnisstatistiken sind in der Tabelle 1 abgetragen (frei nach Döring & Bortz, 2016; Fahrmeir et al., 2003; Stevens, 1946).

Anmerkung: Das Skalenniveau sagt nichts darüber aus, ob eine Skala abzählbar oder nicht abzählbar, also kategorial oder kontinuierlich, respektive diskret oder stetig ist (vgl. Fahrmeier et al., 2003). Nur für Nominalskalen kann klar festgehalten werden, dass diese immer kategorial sind.

Am Beispiel des aggressiven Verhaltens wäre es also möglich mehrere Beobachtungen von Ausprägungen aggressiven Verhaltens (»Schlagen«, »Treten«, »Beleidigen«) mittels unterschiedlicher Niveaus zu beobachten:

- Nominalskala: *trat in einer Stunde auf* (1) oder *trat in einer Stunde nicht auf* (0)

- Ordinalskala: *trat in einer Stunde nicht auf* (0) oder *trat in einer Stunde ein bis zwei Mal auf* (1) oder *trat in einer Stunde mehr als zwei Mal auf* (2)
- Intervall-, Ratio- oder Verhältnisskala: *trat in einer Stunde X mal auf* (x)

3 Testtheorie

Meist wird der Begriff Test für einen standardisierten, strukturierten Test verwendet, welcher aus mehreren standardisierten Items besteht, die normiert und psychometrisch nach den Gütekriterien geprüft sind (Gebhardt, Jungjohann & Schurig, 2021).

In der Testtheorie ist die Basis beispielsweise das gesamte Antwortverhalten von Personen auf Gruppen von Items. Es wird angenommen, dass dies mehreren wiederholten Beobachtungen des gleichen Sachgegenstandes aus unterschiedlichen Perspektiven entspricht. Es sollen Messwerte zugeordnet werden, um damit quantitative Gesetzmäßigkeiten zu formulieren (Steyer & Eid, 2001). Die bedeutsamsten Theorien sind die klassische Testtheorie (KTT) und die probabilistische Testtheorie zu der die Item-Response-Theorie (IRT; Rasch, 1980) gehört. Strenggenommen lassen sich die Angemessenheit der Repräsentation eines Persönlichkeitsmerkmals und die geforderte Eindeutigkeit der Messung, in den Sozialwissenschaften häufig nicht vollständig bestimmen (Diekmann, 2009). Ob die Bestimmung einer Kompetenz perfekt eindeutig und repräsentativ ist kann letztlich nicht bewiesen sondern nur angenommen werden (z. B. Döring & Bortz, 2016, S. 244; Moosbrugger & Kelava, 2008, S. 18f). Die Skalenniveaus und die Bildung der Skalenwerte, welche interpretiert werden sollen (Skalierung), werden nach Konvention festgelegt (z. B. Diekmann, 2009; Döring & Bortz, 2016). Üblicherweise werden diese Eigenschaften einer Messung post hoc anhand der Annahmen gegenüber dem angestrebten Skalenniveau geprüft. Demnach ist auch die Angemessenheit des gewählten Testmodells a priori (vorher) unter Rücksicht auf den Messgegenstand zu planen und post hoc (nach der Messung) zu prüfen. Dafür hat der **Messfehler** eine besondere Relevanz. In allen hier relevanten Testtheorien wird angenommen, dass eine Messung und eine darauf aufbauende Skalierung nicht fehlerfrei sind. Es handelt sich bei der Messung und Skalierung hingegen um eine pragmatische Vereinfachung (Schurig & Kasper, 2018). Für statistische Analysen bedeutet diese Pragmatik eine Reduktion der Realität (zum Beispiel eine schwache arithmetische Kompetenz) in eine formale Struktur (3 von 10 Items gelöst entspricht einem Kompetenzwert von 0.3 bei einem Minimum von 0 und einem Maximum von 1). Diese Annahme ist, damit sie prüfbar wird, mit der Annahme verknüpft, dass Messfehler existieren.

4 Messfehler

Der bekannteste Messfehler ist dabei der Standardfehler des Mittelwerts (SE; $\sigma_{\bar{x}}$). Dieser wird aus der Standardabweichung σ eines Mittelwerts (Vorraussetzung: Intervallskala) bestimmt als

$$\sigma_{\bar{x}} = (\frac{\sigma}{\sqrt{n}}) \tag{2}$$

Die Standardabweichung einer Stichprobe wird bestimmt als

Tabelle 2: Exemplarische Ergebnismatrix

	Item 1	**Item 2**	**Item 3**	**Item 4**	**Item 5**	**Item 6**	**Item 7**	**Item 8**	**Item 9**	**Item 10**	$\sum$
Person 1	0	0	1	0	1	0	0	1	0	0	3
Person 2	0	1	1	1	0	0	1	0	1	0	5
Person 3	0	0	0	0	0	1	1	0	0	1	3
Person 4	1	1	1	1	1	1	1	1	1	0	9
Person 5	1	1	1	1	0	1	1	1	1	1	9
Person 6	1	1	1	1	1	0	1	1	1	1	9
Person 7	1	0	0	1	1	1	0	0	1	1	6
Person 8	1	1	1	1	1	1	1	1	1	1	10
Person 9	1	1	0	0	1	1	1	1	1	1	8
Person 10	1	1	1	1	1	1	0	1	0	1	8
Richtig gelöste Items	7	7	7	7	7	7	7	7	7	7	–
Mittelwert	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	7
SD	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48	2.58
SE des Mittelwerts	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.82

$$\sigma = \sqrt{\frac{(Wert_1 - Mittelwert)^2 + (Wert_{...} - Mittelwert)^2 + (Wert_i - Mittelwert)^2}{n-1}} \tag{3}$$

Die Standardabweichung ist ein Maß für die Streubreite der Werte rund um das arithmetische Mittel dieser Werte. Anders gesprochen ist die Standardabweichung die durchschnittliche Abweichung aller Werte vom Mittelwert. Die Maßeinheit der Standardabweichung entspricht also der des gemessenen Wertes. Der Standardfehler des Mittelwerts ist wiederum ein Maß dafür, wie gut der Mittelwert einer Stichprobe einen (theoretischen) Populationsmittelwert schätzt. Der so errechnete Messfehler (in diesem Beispiel der Messfehler eines Mittelwertes) gibt an, wie eng die Werte aneinander liegen und kann verwendet werden um die Angemessenheit einer Messung und eines Tests über die Hauptgütekriterien von Tests zu belegen (vgl. Grabowski, Castello & Brodersen, in diesem Band). Je kleiner ein Messfehler, desto präziser ist ein Wert.

Wenn alle 10 Items des Mathetests eine gleiche mittlere Lösungswahrscheinlichkeit von .7 aufweisen, so könnte eine Ergebnismatrix wie in Tabelle 2 dargestellt aussehen.

Die Berücksichtigung von Messfehlern wird überdisziplinär als notwendig erachtet, wenn in Analysen generalisiert, also verallgemeinert, werden soll. Relevant ist dabei zu beachten, dass der Messfehler nicht nur für Skalenwerte, sondern auch für einzelne Items (über Personen hinweg) oder auch innerhalb von Personen (über Items) bestimmt werden kann (Rost, 2004).

Es wird davon ausgegangen, dass es drei zentrale Bedrohungen des Messfehlers gibt. Zufällige Fehler, die nicht kontrollierbar sind, wie zum Beispiel schlechtes Wetter am Testtag. Hier kann davon ausgegangen werden, dass sich die Fehler bei mehreren Messungen (unterschiedliche Testteilnehmer) ausgleichen. Fahrlässige Fehler beruhen auf menschlichem Verhalten, wie zum Beispiel unabsichtlichem »Vorsagen« des/der Testleiter:in. Zuletzt sind systematische Fehler

die größte Bedrohung. Diese umfassen unscharfe Messinstrumente, welche durch einen hohen Messfehler erkannt werden können. Wie aber bestimmt und interpretiert man Messfehler? Dies hängt von der jeweiligen Testtheorie ab.

5 Konzepte von Messfehlern in KTT und IRT

5.1 Klassische Testtheorie

Die KTT stellt eine Sammlung von Methoden zur Erfassung interindividueller Unterschiede dar (Lord, Novick & Birnbaum, 1968). Diese Methoden werden als klassisch bezeichnet, da es neben diesen Methoden auch andere – weniger etablierte – Alternativen gibt (Rost, 2004). Zentral wird angenommen, dass sich das beobachtete Merkmal aus dem »echten« Merkmal (*T*; true score), also beispielsweise der Kompetenz, und einem zufälligen Messfehler (*e*; error). Die Messfehler hängen dabei (so die theoretische Annahme) nicht mit *T* oder anderen Werten zusammen. Die additive Verbindung ergibt den Testwert ($T + e = X$). Die KTT geht davon aus, dass (*T* innerhalb eines Individuums konstant ist und *e* sich bei wiederholten Messungen herausmitteln würden, da dieser zufällig ist. Würde also bei einer Person theoretisch unendlich oft wiederholt eine Fähigkeit gemessen, entspräche der Mittelwert dieser Messungen dem wahren Fähigkeitswert. Daher werden zur Erfassung von *X* bei einer Person mehrere Items verwendet. Es wird angenommen, dass diese die gleiche (interessierende) Fähigkeit adressieren. Jedes dieser Items wird dabei als wiederholte Messung aufgefasst. Messfehler werden dabei durch die Zusammenfassung der einzelnen Messungen zu einem Testwert ausgeglichen. Wenn also die arithmetische Kompetenz einer Person 10-mal gemessen würde (also mit 10 Wiederholungen des gleichen Instruments oder 10 Items innerhalb eines Instruments), so würde angenommen werden, dass in jeder der Messungen ein zufälliger Messfehler enthalten ist. Da der erwartete Mittelwert des Messfehlers in der KTT 0 ist, gilt entspricht ($T = X$. Der beobachtete Testwert, z. B. ein Wert von 5 für Person 2 im Datenbeispiel, ist also gleich dem zu beobachtenden »echten« Merkmal *T*.

Dies bedeutet, dass die KTT besonders empfindlich für systematische Messfehler ist. Dieser Messfehler geht bei jeder wiederholten Messung mit ein. Wenn also alle 10 Items oder Wiederholungen die sprachlichen Fähigkeiten mit einbeziehen, weil die Aufgaben in einem Textformat präsentiert werden, dann liegt ein systematischer Messfehler vor. Schüler:innen mit hoher sprachlicher Kompetenz können die arithmetischen Aufgaben besser lösen als die anderen Kinder und der Test wäre nicht fair. Der Score des Tests der eigentlich nur die arithmetische Kompetenz beschreiben soll, würde durch eine andere Fähigkeit konfundiert (vermischt).

Aus den Annahmen der KTT lassen sich Aussagen über die Reliabilität und weitere Eigenschaften eines Tests (Gütekriterien) ableiten (Moosbrugger, 2008; Grabowski, Castello & Brodersen, in diesem Band). Die Vorzüge der KTT liegen in der Einfachheit der zugrundeliegenden Annahmen sowie der daraus resultierenden leichten empirischen Realisierbarkeit. So kann für Tests und Fragebögen zur Prüfung der internen Konsistenz der einfach her leitbare Cronbach's Alpha (α) als Standard verwendet werden. Der α beschreibt das mittlere Maß des Zusammenhangs zwischen den verwendeten Items (d.h. wie gut die Items das Konstrukt im Mittel messen). Die zentrale Annahme dabei ist, dass alle Items gleichermaßen und relativ gleichförmig (Äquivalenz der Faktorladungen oder tau-Äquivalenz; vgl. Moosbrugger & Kelava, 2008) das interessierende Merkmal abbilden. Der α wird bestimmt aus der Anzahl der verwendeten Beobachtungen (Items; *T*) und dem mittleren Zusammenhang (Korrelation) der Beobachtungen ($\bar{r}$).

Tabelle 3: COTAN System zur Einschätzung von Reliabilitätswerten (Evers et al. 2019)

	Test für wichtige Entscheidungen auf individuellem Niveau (z. B. Feststellungsverfahren für Förderbedarf)	Test für weniger wichtige Entscheidungen auf individuellem Niveau (z. B. Fördererfolg, Verlaufsdiagnostik)	Test auf Gruppenniveau (z. B. Klassenklima)
Gut	>.9	>.8	>.7
Genügend	.8 – .9	.7 – .8	.6 – .7
Ungenügend	<.8	<.7	<.6

$$\alpha = \frac{N * \bar{r}}{1 + (N-1) * \bar{r}} \tag{4}$$

Für dichotome Items (mit zwei Ausprägung wie falsch »0« und richtig »1«) wie in der Tabelle 1 wird die Kuder-Richardson Formel (es gibt verschiedene Versionen) verwendet, welche sich als

$$\alpha_{KR20} = \frac{1}{N-1} * (1 - \frac{Mittelwert * (N - Mittelwert)}{N * \sigma^2}) \tag{5}$$

zusammensetzt. Damit ergibt sich für die Werte in Tabelle 1 eine interne Konsistenz von $1.11 * (1 - \frac{7*(10-7)}{10*6.67}) = 0.76$. Viele Statistikprogramme (z. B. SPSS) verwenden automatisch die Kuder-Richardson Formel, wenn der α errechnet wird. Das *R* Paket psych verwendet hingegen den α als Generalisierung. Die Höhe der internen Konsistenz hängt also vom Verhältnis der Beobachtungen zueinander, zum Gesamtwert und der Anzahl der Items ab. Die Höhe kann entsprechend der Empfehlungen des Niederländischen Psychologenverbandes (Nederlands Instituut van Psychologen, NIP) als angemessen angesehen werden, wenn keine relevanten individuellen Entscheidungen (z. B. Feststellungsverfahren) damit getroffen würden (Tabelle 3; Evers et al., 2019). Oft werden durch Fachverbände keine konkreten Schwellenwerte angegeben, da die Schätzung der Angemessenheit der Reliabilität eines Tests stark vom Einsatzgebiet und der notwendigen Präzision des Testwerts abhängt. Eine dezidierte Beschreibung der Schwierigkeit bei der Einschätzung angemessener Schwellenwerte findet sich bei Schermelleh-Engel und Werner (2012).

Der Nachteil ist, dass sich die Axiome der KTT in ihrem eigenen Rahmen kaum prüfen lassen, da sowohl der wahre Wert als auch der Fehler nicht direkt beobachtbar sind (Steyer & Eid, 2001). Somit müsste für eine Prüfung die zentrale Annahme der unzusammenhängenden Fehler gelöst werden. Die Annahme ist also oft zu grob. Eine Option zur Prüfung stellen Faktoranalysen da, welche die Annahmen, insbesondere zur Unabhängigkeit der Fehler, lockern. Zur adäquateren Reliabilitätsabschätzung existieren aber auch weitere unterschiedlich ausdifferenzierte Formeln (z. B. zur Lockerung der Annahme der gleichen Faktorladungen kann die essentielle tau-Äquivalenz über den Omega; Ω bestimmt werden; vgl. Steyer & Eid, 2001). Der Ω gibt üblicherweise etwas höhere Reliabilitäten aus, da weniger zu prüfende Annahmen gemacht werden, zugleich gibt es aber noch weniger Erfahrungen und empfohlene Schwellenwerte zum Umgang mit dem Wert. Der Ω wird beispielsweise standardmäßig bei Reliabilitätsanalysen mit dem *R* Paket psych mit ausgegeben. Auch eine Ausgabe mit SPSS ist möglich.

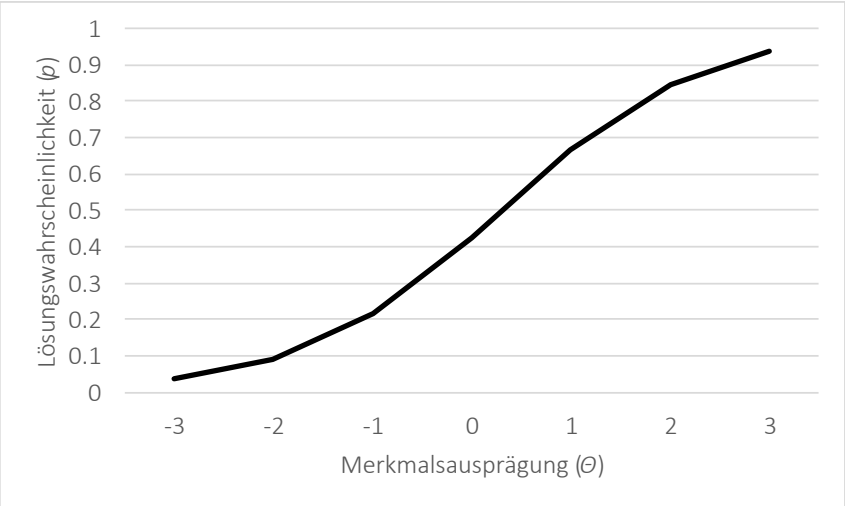

Abbildung 1: Item Characteristic Curve (ICC): Wahrscheinlichkeit ein Item (*p*) zu lösen bei gegebener Merkmalsausprägung (θ)

Eine zentrale Annahme ist aber, dass die Eigenschaften der Items nur für die vorliegende Stichprobe Gültigkeit haben. Dies umfasst die Reliabilität, die Itemschwierigkeiten, die Itemtrennschärfen und weitere Parameter auf Itemebene.

5.2 Item-Response-Theorie

Angesichts dieser und weiterer Einwände ist die probabilistische Testtheorie (Item-Response-Theorie; IRT; Rost, 2004) als wichtige Erweiterung und Ergänzung zur KTT zu sehen. Zentral ist in der IRT die Modellgleichung zur Bestimmung der Lösung eines Items x, wenn der Persönlichkeitsparameter θ und die Schwierigkeit das Item zu lösen σ_1 gegeben ist. Die Schwierigkeit das Item zu lösen ist im einfachsten Fall von dichotomen Items, bei gegebener Eichstichprobe, gleich dem Prozentsatz der in der Stichprobe gelösten Items.

$$p(x = 1) = \frac{exp(\theta - \sigma_1)}{1 + exp(\theta - \sigma_1)} \tag{6}$$

Für die Werte in Tabelle 1 würde dies heißen, dass alle Items eine Schwierigkeit von $1 - 0.7 = 0.3$ haben. Wenn θ auf einer Metrik von -3 bis 3 abgetragen würde, so würde sich also ergeben, dass die Wahrscheinlichkeit eines der Items zu lösen bei $\theta = -2$ gleich 0.09, also 9% wäre $\frac{exp(-2-0.3)}{1+exp(-2-0.3)}$.

IRT richtet sich vor allem an abzubildende inhaltliche Bereiche, die eine breite Spannweite von Merkmalsausprägungen umfassen, wie zum Beispiel in Kompetenztests. Daher ist es für eine Erstellung eines Erhebungsinstruments im Sinne der IRT unpraktisch, wenn alle Items die gleichen Lösungswahrscheinlichkeiten pro Merkmalsausprägung haben. Üblicher ist, dass eine Spannweite von unterschiedlich schweren Aufgaben gewählt wird. Items sind also üblicherweise für unterschiedliche Schwierigkeitslagen unterschiedlich trennscharf. Am häufigsten werden die meisten Items im mittleren Schwierigkeitsbereich angelegt, da auch die meisten Schüler:innen auch im mittleren Fähigkeitsbereich verortet sind (siehe Renner, in diesem Band). Es ist aber mittels der theoretischen Fundierung und der Inspektion und Eichung der Items (nicht des Tests,

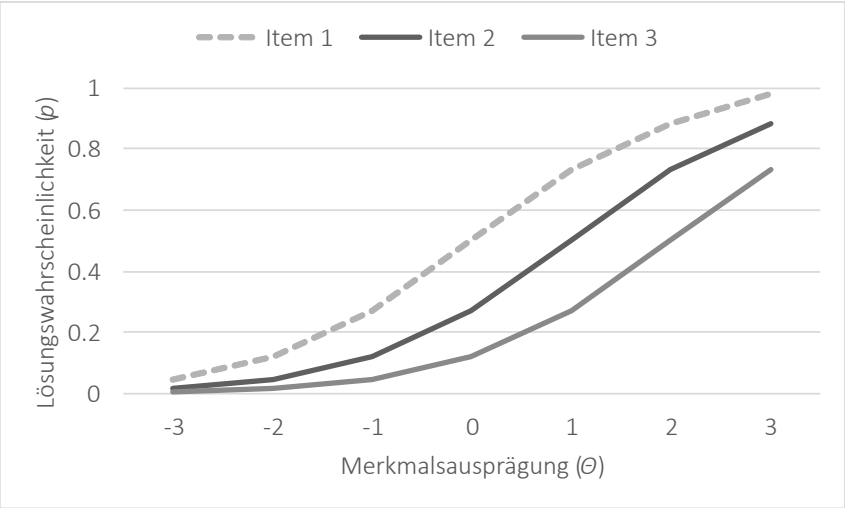

Abbildung 2: Item Characteristic Curves (ICC): Kombinierte Wahrscheinlichkeiten mehrere Items zu lösen (*p*) bei gegebener Merkmalsausprägung (θ)

dieser Ansatz würde in der KTT verfolgt!) möglich, Gesamttests so anzulegen, dass diese besonders leicht oder besonders schwer sind, also in spezifischen Fähigkeitslagen besonders gut differenzieren.

Relevant für die Feststellung der Reliabilität von IRT Modellen ist die Feststellung der Präzision, mit der ein Testwert ermittelt wird. Diese differiert dabei je nach der Lage des Merkmals im beobachteten Leistungskontinuum (also dem Wertbereich von θ), die Genauigkeit der Messung ist im mittleren Bereich üblicherweise besser als im extremen Bereich (Rost, 2004), da in diesem Bereich mehr Items eingesetzt werden. Aus den Ungenauigkeiten am Rand des Testbereiches kommt es zu erhöhten Standardfehlern in den Extrembereichen, da diese durch die häufig gegebene Normalverteilung der Personenparameter geringer besetzt sind. Dies erschwert die Schätzung einer absoluten Messgenauigkeit im Kontext der IRT ohne weitere Annahmen.

Aufgrund der vorgestellten Eigenschaften der Varianzverteilungen in IRT Modellen ist es üblich, anstatt eines absoluten Maßes der Reliabilität lokale Item- und Personenparameter zu prüfen, um auf die Reliabilität zu schließen. Daher steht im Kontext der IRT immer die Passung der einzelnen Items zum Modell sowie zum Merkmal und der adressierten Schwierigkeitslage im Vordergrund (Steyer & Eid, 2001; Rost, 2004).

6 Prüfung und Gütekriterien von Tests

Um die Güte der Repräsentation einzuschätzen, müssen Tests formal identifizierbar und testbar sein (Steyer, 1989). Dabei wird heute nur noch selten davon ausgegangen, dass ein Modell absolut oder vollständig zu den Daten passt oder nicht passt. Es steht nun vielmehr die Güte der Anpassung im Vordergrund, welche durch sogenannte Fit-Indizes beschrieben werden, die die Anpassung des Modells an die Daten beschreiben (z. B. Kline, 2011, S. 190). Während die Probleme der Identifizierbarkeit und Testbarkeit zentral in der probabilistischen Testtheorie behandelt wurden, zeigt Steyer (1989) auf, dass die Fragen auch für die Modelle der klassischen Testtheorie relevant sind. Üblicherweise werden dabei mehrere Strategien verfolgt. Zum einen kann die Anpassung des theoretischen Modells an die gegebenen Daten sowie lokale Parameter geprüft werden. Zum anderen können alternative testtheoretische Modelle (z. B. gegen ein

sogenanntes Nullmodell) geprüft werden. Für eine Einführung zu Strategien der Modellbeurteilung vgl. Schurig (2017, S. 80).

7 Fazit

Die Güte der verwendeten statistischen Modelle und die damit verbundene Beachtung und Prüfung derer zugrunde liegenden Annahmen befördert wiederum die Tragkraft der daraus abgeleiteten Aussage! Dies wurde eindrucksvoll von Reynolds (2010) beschrieben, nachdem zum Herausgeber der Fachzeitschrift *Psychological Assessment* gewählt wurde. Die Ergebnisse von richtig konzipierten Arbeiten haben eine größere Wahrscheinlichkeit, die die Forschung voranzubringen und die Praxis zu verbessern. Und als Wissenschaftler sollten wir den Verlust von guten Ideen aufgrund der schlechten Ausführung vermeiden.

Tests haben theoretische Annahmen und messen die Empirie anhand formaler, mathematischer Regeln. Daher werden Antworten oder Reaktionen eindeutige Zahlenwerte zugeordnet, die ein latentes Konstrukt möglichst genau repräsentieren. Die Zahlen werden in ein mathematisches Mess- oder Testmodell überführt, wodurch durch Prüfungen auf Modellgültigkeit und eine Prüfung der Modellparameter Widersprüche im formalen Modell oder gegebenenfalls in der Theorie ausgemacht und Prognosen abgeleitet werden (Schurig, 2017). Die Akzeptanz der vorgestellten imperfekten Repräsentation und der häufig unvollständigen Anpassung statistischer Modelle an die Daten hat einen direkten Effekt auf die praktische Betrachtungsweise, welcher abgeleitete Zusammenhänge unterliegen müssen. Diese Akzeptanz reflektiert sich im Umgang mit empirischen Sätzen und deren Abgrenzung zu normativen Sätzen (Popper, 1994, S. 7ff) und dem daraus erwachsenden Verständnis von interessierenden substanzwissenschaftlichen Wirkungszusammenhängen.

Die Annahmen zu den Eigenschaften der diesen Annahmen zugrundeliegenden (Mess-)Fehler sind an fachspezifische Interpretationskulturen geknüpft. Wenn zum Beispiel ganze pädagogische Prozesse betrachtet werden sollen, müssen Annahmen zur Zufälligkeit und Zusammenhangslosigkeit von Fehlern oft zwangsweise häufig zurückgewiesen werden, weil beispielsweise pädagogische Prozesse in Lerngruppen und strukturiert verlaufen (Schurig & Kasper, 2018). Für die sonderpädagogische Diagnostik sind Tests und Fragebögen ebenso immer im Kontext zu interpretieren. So wird es in der Sonderpädagogik immer auch kleinere Stichproben und Einzelfalldiagnostik geben, da sich die Sonderpädagogik mit speziellen Fällen sowie pädagogischen Herausforderungen und nicht mit dem Durchschnitt der Gesellschaft beschäftigt. Insbesondere angesichts der Komplexitätsgrade von Prozessen zwischen Individuen, pädagogischen Akteuren, Systemen und Lebenswelten drängt es sich auf, dass (Mess-)Fehler stärker interpretiert und analysiert werden, um möglichst valide Evidenz (Newton & Shaw, 2013) zu erzeugen.

Literatur

Diekmann, A. (2009). *Empirische Sozialforschung: Grundlagen, Methoden, Anwendungen* (20. Aufl.). Rowohlt Taschenbuch Verlag.

Döring, N. & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften.* Springer. https://doi.org/10.1007/978-3-642-41089-5

Evers, A. V. A. M., Lucassen, W., Meijer, R. R., & Sijtsma, K. (2019). *COTAN assessment system for quality of tests.* https://www.psynip.nl/wp-content/uploads/2019/05/NIP-Brochure-Cotan-2018-correctie-1.pdf

Fahrmeir, L., Künstler, R., Pigeot, I. & Tutz, G. (2003). *Statistik: Der Weg zur Datenanalyse.* Springer. https://doi.org/10.1007/978-3-662-22657-5

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests: Grundlagen und Anwendungen*. Huber.

Gebhardt, M., Jungjohann, J. & Schurig, M. (2021). *Lernverlaufsdiagnostik im förderorientierten Unterricht: Testkonstruktionen, Instrumente, Praxis.* Ernst Reinhardt Verlag.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3. ed.). Guilford Press.

Lienert, G. A. & Raatz, U. (1998). *Testaufbau und Testanalyse* (6. Aufl.). Beltz.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39* (2), 247–264. https://doi.org/10.1007/BF02291471

Lord, F.M., Novick, M.R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Lüdtke, O., Robitzsch, A., Trautwein, U., & Köller, O. (2007). Umgang mit fehlenden Werten in der psychologischen Forschung: Probleme und Lösungen. *Psychologische Rundschau, 58* (2), 103–117. https://doi.org/10.1026/0033-3042.58.2.103

Moosbrugger, H. & Kelava, A. (Hrsg.). (2012). *Testtheorie und Fragebogenkonstruktion*. Springer. https://doi.org/10.1007/978-3-642-20072-4

Newton, P. E., & Shaw, S. D. (2013). Standards for talking and thinking about validity. *Psychological Methods, 18* (3), 301–319. https://doi.org/10.1037/a0032969

OECD (2022), *Poverty rate (indicator)* . https://doi.org/10.1787/0fe1315d-en (12 January 2022)

Orth, B. (1974). *Einführung in die Theorie des Messens*. Kohlhammer.

Popper, K. R. (1994). *Logik der Forschung* (10. Aufl.). Mohr.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Univ. of Chicago Press.

Reynolds, C. R. (2010). Measurement and assessment: An editorial view. *Psychological Assessment, 22* (1), 1–4. https://doi.org/10.1037/a0018811

Rost, J. (2004). *Lehrbuch Testtheorie, Testkonstruktion* (2. Aufl.). Huber.

Schermelleh-Engel K., Werner C. S. (2012). *Methoden der Reliabilitätsbestimmung*. In: Moosbrugger H., Kelava A. (Hrsg.) Testtheorie und Fragebogenkonstruktion. Springer. https://doi.org/10.1007/978-3-642-20072-4_6

Schurig, M. (2017). *Latente Variablenmodelle in der empirischen Bildungsforschung – Die Schärfe und Struktur der Schatten an der Wand* [Dissertation]. TU Dortmund, Dortmund. https://doi.org/10.17877/DE290R-18044

Schurig, M. & Kasper, D. (2018). Zur Interpretation von Messfehlern aus Sicht der Erziehungswissenschaft. *Erziehungswissenschaft*, 56, 45–54. https://doi.org/10.3224/ezw.v29i1.01

Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science, 103* (2684), 677-680.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review, 64* (3), 153–181. https://doi.org/10.1037/h0046162

Steyer, R. (1989). Models of Classical Psychometric Test Theory as Stochastic Measurement Models: Representation, Uniqueness, Meaningfulness, Identifiability, and Testability. *Methodika*. (3), 25–60.

Steyer, R. & Eid, M. (2001). *Messen und Testen: Mit Übungen und Lösungen*. Springer. https://doi.org/10.1007/978-3-642-56924-1

Vertr. Prof. Dr. Michael Schurig ist Erziehungswissenschaftler mit einem Schwerpunkt in Forschungsmethoden, deren Anwendung und inklusiver Schulentwicklung. Er ist Akademischer Oberrat und vertritt zurzeit die Professur für die Entwicklung und Erforschung inklusiver Bildungsprozesse an der Fakultät Rehabilitationspädagogik der TU Dortmund. https://orcid.org/0000-0002-7708-0593

Prof. Dr. Markus Gebhardt ist Sonderpädagoge und Lehrstuhlinhaber für Lernbehindertenpädagogik einschließlich inklusiver Pädagogik an der Universität Regensburg. https://orcid.org/0000-0002-9122-0556