

Grundkonzepte der Skalierung

Basic Concepts of Scaling

Thomas Staufenbiel & Ingwer Borg

1 Skalierung als Zuweisung von Messwerten

Im Rahmen psychologischer Beobachtungen werden verschiedene Instrumente eingesetzt (z. B. ein Intelligenztest, eine Stoppuhr oder „die Lehrerin“), die Messwerte produzieren. Aus dem Intelligenztest resultiert für den Bewerber ein Intelligenzquotient (IQ) von „120“, die Stoppuhr registriert eine Laufzeit der Ratte von „92“ Sekunden und die Klassenarbeit eines Schülers wird durch die Lehrerin mit der Schulnote „2“ bewertet.

Begriffsklärung:

Die Zuweisung von Zahlen, die die Ausprägung in einem interessierenden Merkmal ausdrückt, wird häufig als *Messen* oder *Skalierung* bezeichnet. Stevens (1959) definiert Messen entsprechend als „... Zuweisung von Zahlen zu Objekten oder Ereignissen gemäß einer Regel ...“ (S. 18).

Die Regel, nach der die Zuweisung erfolgt, kann einfach sein, wie im Beispiel der Messung der Laufgeschwindigkeit der Ratte, in der die Sekunden auf der Stoppuhr abgelesen werden. Sie kann aber auch, wie bei der Intelligenzmessung, komplizierter sein: So kann man zunächst den Testaufgaben, die richtig beantwortet werden, den Wert 1 und denen, die falsch beantwortet werden, den Wert 0 zuweisen; die Skalenwerte dann über die Items aufsummieren; und diese Summenwerte schließlich auf eine Skala transformieren, die den Mittelwert 100 und die Standardabweichung 15 hat (wie dies für die IQ-Skala gilt).

Die Zuweisung der Zahlen beim Skalieren soll jedoch nicht nur nach irgendeiner willkürlichen Regel erfolgen, sondern so, dass man anhand der Zahlen *bedeutungsvolle* Aussagen über den Beobachtungsgegenstand treffen kann. So sollte ein Bewerber mit dem Wert 120 auf der Intelligenzskala intelligenter sein als einer mit einem IQ von 105 und ein Schüler mit einer „2“ in einer Klassenarbeit die Leistungsanforderungen besser beherrschen als einer mit einer „3“ auf der üblichen Notenskala. Darüber hinaus soll sich möglicherweise auch ein Notendurchschnitt ergeben, der valide Aussagen über die allgemeine Leistung des Schülers erlaubt. Die Frage, welche Aussagen man auf der Grundlage von Skalenwerten treffen kann, wird unter dem Begriff *Skalenniveau* thematisiert.

2 Skalenniveau

Das Skalenniveau bezeichnet die Transformierbarkeit einer Skala s in eine andere Skala s' . Die Messung der Laufzeit einer Ratte auf der Zeitskala in Sekunden weist einen natürlichen Nullpunkt auf; der Wert 0 steht für keine verstrichene Zeit oder Gleichzeitigkeit. Dies gilt ebenso für andere physikalische Skalen wie Gewicht oder Geschwindigkeit. Eine Verschiebung der Skalenwerte durch die Addition einer Konstanten d ist daher nicht sinnvoll. Möglich ist aber, die Skalenwerte beispielsweise in Minuten zu transformieren, ohne dass dadurch irgendetwas von Bedeutung verloren geht. Eine solche Maßstabsänderung entspricht der Multiplikation mit der Konstanten $c = 1/60$. Skalen, bei denen die Skalenwerte nur dadurch verändert werden dürfen, dass man sie mit einer Konstanten $c > 0$ multipliziert, bezeichnet man als *Verhältnisskalen*. Welche Beziehungen zwischen den Skalenwerten sind nun deutbar? Allgemein sind dies alle, die sich bei *zulässigen Transformationen* nicht ändern. Bei Verhältnisskalen sind dies die Verhältnisse zwischen Skalenwerten. Eine Ratte a , die das Labyrinth in $s_a = 90$ Sekunden durchläuft, braucht *halb* so lange wie eine andere Ratte b , die dafür $s_b = 180$ Sekunden benötigt. Transformiert man die Sekundenskala mittels $s'_x = s_x/60$ in eine Minutenskala, so bleibt das Verhältnis $s_a/s_b = 90/180 = 1/2$ auf der neuen Skala erhalten, da $s'_a/s'_b = 1.5/3 = 1/2$. Aussagen über Verhältnisse ändern sich unter Transformationen der Art $s'_x = c \cdot s_x$ mit $c > 0$ nicht. Die für Verhältnisskalen erforderliche Annahme eines bedeutsamen Nullpunktes lässt sich jedoch für psychologische Merkmale (wie z. B. Null Intelligenz) kaum begründen. Daher lassen sich psychologische Merkmale nur selten (falls überhaupt) auf Verhältnisskalenniveau messen.

Nimmt man an, dass Messwerte auf einer *Intervallskala* liegen, so ist neben der Multiplikation mit einer Konstanten auch die Addition einer Konstanten zulässig: $s'_x = c \cdot s_x + d$, mit $c > 0$ (*lineare Transformation*). Hier sind nun Aussagen über Verhältnisse – z. B. eine Person mit einem Wert von $s_a = 140$ in einem Intelligenztest sei doppelt so intelligent wie eine andere mit einem Wert von $s_b = 70$ – nicht mehr zulässig, weil $s_x/s_y \neq s'_x/s'_y = [c \cdot s_x + d]/[c \cdot s_y + d]$, falls $d \neq 0$. Erhalten bleibt auf der Intervallskala hingegen das *Verhältnis von Differenzen* unter allen linearen Transformationen. Beträgt der *Intelligenzunterschied* zwischen zwei Personen u Intelligenzpunkte (z. B. wie oben $u = 140 - 70 = 70$) und der zwischen zwei anderen Personen v Intelligenzpunkte, so ändert sich das Verhältnis u/v unter linearen Transformationen nicht. Weisen die beiden anderen Personen z. B. Skalenwerte von $s_c = 90$ und $s_d = 125$ auf, so ist $v = 35$ und damit bei beliebigen linearen Transformationen der Skalenwerte der vier Personen der *Intelligenzunterschied* von c und d immer halb so groß wie der von a und b .

Wenn man davon ausgeht, dass die Werte einer Skala nur in Bezug auf ihre Rangordnung bedeutsam sind (also ob eine Merkmalsausprägung stärker als eine andere ist oder ob beide gleich sind), dann handelt es sich um eine *Ordinalskala*. Für

sie sind alle Transformationen zulässig, die zu Skalenwerten führen, die genauso geordnet sind wie die Ausgangswerte. Der Abstand der Objekte auf einer Ordinalskala ist dagegen ohne Bedeutung und kann nicht interpretiert werden. Dies gilt beispielsweise für die Variable Schulabschluss: Dort sind die Ausprägungen 3=Gymnasium, 2=Realschule und 1=Hauptschule geordnet im Sinne des Bildungsniveaus; sie bilden eine Ordinalskala. Verhältnisse und Differenzen kann man für diese Skalenwerte zwar berechnen, aber sie haben keine offensichtliche inhaltliche Bedeutung. Personen mit einem Hauptschulabschluss sind jedenfalls nicht „doppelt so gebildet“ wie Personen mit Realschulabschluss.

Das schwächste Skalenniveau weist die *Nominalskala auf*, die man oft bei einfachen Codierungen findet (z. B. Geschlecht in 1 = männlich, 2 = weiblich). Hier ist nur noch von Bedeutung, ob zwei Objekte gleich (*äquivalent*) oder verschieden sind. Alle Personen mit dem Skalenwert 2 sind Frauen, alle mit dem Skalenwert 1 sind Männer. Dass der zur Codierung verwendete Wert 2 größer ist als der Wert 1, ist unerheblich. Entsprechend sind alle Transformationen zulässig, die Verschiedenheit und Äquivalenz der Skalenwerte der Objekte erhalten. Man könnte also Männer z. B. auch mit 23 codieren und Frauen mit 3.

Die Skalenniveaus sind untereinander *geordnet*. Das Nominalskalenniveau ist das schwächste Niveau in dem Sinn, dass die größten Freiheiten für die Transformationen der Skalenwerte bestehen. Die Nominalskala enthält daher auch nur wenig Information: Sie zeigt nur die Verschiedenheit und die Gleichheit von Objekten an. Die Verhältnisskala ist die stärkste Skala: Aus ihr kann man nicht nur auf Gleichheit und Verschiedenheit der Objekte schließen, sondern man kann auch das Verhältnis von Skalenwertdifferenzen als auch von Skalenwerten selbst inhaltlich deuten.

Merke:

Transformationen, die auf einem Skalenniveau zulässig sind, dürfen immer auch bei allen schwächeren angewendet werden. Umgekehrt sind Schlussfolgerungen, die auf einem Skalenniveau zulässig sind, auch auf allen stärkeren möglich.

In der Praxis ist die Feststellung, welches Skalenniveau bestimmte Messwerte aufweisen, nicht leicht zu treffen. Exemplarisch wird dies an Schulnoten deutlich, für die häufig Intervallskalenniveau reklamiert wird. Andere sind dagegen der Auffassung, dass Schulnoten allenfalls als ordinal deutbar sind, weil man nicht annehmen könne, dass der Unterschied in der Leistung zwischen Schülern mit den Noten 1 und 2 genauso groß ist, wie der zwischen Schülern mit den Noten 2 und 3 (oder 3 und 4): Vielmehr könne man nur sagen, dass ein Schüler mit einer besseren Note eine bessere Leistung gezeigt hat. Die Begründung des Skalenniveaus von Mess-

werten sollte also darauf aufbauen, dass man irgendwie (argumentativ oder empirisch) nachweist, dass Aussagen über Ordnung, Differenzen oder Verhältnisse der Skalenwerte in sich konsistent und inhaltlich bedeutsam sind.

Velleman und Wilkinson (1994) schlagen dagegen vor, das Skalenniveau von Daten als eine *Rollenzuweisung* zu verstehen, die durch den Untersucher erfolgt. Die Nützlichkeit dieser Zuweisung soll sich durch die Bestätigung falsifizierbarer Vorhersagen (z. B. IQ-Werte korrelieren positiv mit Berufserfolg) empirisch erweisen.

3 Skalenkonstruktion und Skalenanalyse

Ein Teilgebiet der Skalierung beschäftigt sich damit, Messinstrumente (wie z. B. einen Intelligenztest) so zu *konstruieren*, dass ihre Skalenwerte bestimmte wünschenswerte Eigenschaften aufweisen. Bei der Konstruktion spielen verschiedene Überlegungen eine Rolle, wie etwa solche zum Gegenstand der Skalierung (z. B.: Gibt es nur eine Intelligenz oder verschiedene Formen?), zur Formulierung der Items (hier: der Intelligenzaufgaben), zur Methode der Datenerhebung (z. B. Verwendung der Testaufgaben zur direkten Beantwortung oder ihre Vorgabe im Paarvergleich) und zur Verrechnung der Antworten zu einem Skalenwert.

Im Gegensatz zu dieser Vorgehensweise der *Skalenkonstruktion* wird bei der *Skalenanalyse* geprüft, ob gegebene Daten bestimmte Skaleneigenschaften besitzen. Nehmen wir an, wir wollten eine Skala der Kundenorientierung verschiedener Baumärkte erstellen. Dazu befragen wir eine Stichprobe von Heimwerkern. Wir verwenden hierbei ein Messinstrument, bei dem einem Heimwerker immer Paare von Baumärkten vorgegeben werden (z. B. Baumarkt O vs. P, H vs. O usw.) und dieser dann jeweils beurteilen muss, welcher der beiden Baumärkte kundenorientierter ist. Aus diesen Daten aller Befragten wird dann eine gemeinsame Skala erstellt (z. B. mittels des LCJ-Verfahrens, vgl. Abschnitt 4.2). Diese Schritte sind Bestandteil der Skalenkonstruktion. Mittels Skalenanalyse lässt sich dann auf vielfältige Weise prüfen, wie gut diese Skala ist:

1. Zunächst kann man betrachten, ob verschiedene Heimwerker (zumindest annähernd) zu gleichen Urteilen kommen. Wäre dies nicht der Fall, so würde eine allgemeine Kundenorientierungsskala der Baumärkte nicht viel Sinn machen.
2. Man kann die Datenerhebung wiederholen und prüfen, ob wieder ähnliche Urteile abgegeben werden. Resultiert bei einer solchen Re-Testung der gleichen Personen eine ganz andere Skala, so spricht dies gegen die Qualität (vor allem die Stabilität) der Skala.
3. Man kann aufgrund der Skala bestimmte theoretisch abgeleitete Vorhersagen prüfen. Derartige Vorhersagen wären z. B., dass die Baumärkte mit einer besseren Kundenorientierung auch allgemein ein besseres Image aufweisen oder einen größeren Umsatz machen. Treten diese Vorhersagen ein, so spricht dies

für die Kundenorientierungsskala. Würde sie nämlich etwas ganz anderes erfassen oder die Kundenorientierung sehr unzuverlässig messen, so wäre kaum erklärlich, warum sich die vorhergesagten Zusammenhänge ergeben.

4. Man kann überprüfen, ob die Urteile der Heimwerker in sich stimmig sind. Sagt ein Heimwerker beispielsweise, dass er Baumarkt H kundenorientierter beurteilt als Baumarkt O und O wiederum als P, so sollte er beim Vergleich der Paarlinge H und P Baumarkt H als kundenorientierter einschätzen. Ist dies nicht der Fall, so sind die Urteile inkonsistent. Treten solche Inkonsistenzen (auch als *zirkuläre Triaden* bezeichnet) in größerem Ausmaß auf, so sind die Urteile der Heimwerker nicht skalierbar – jedenfalls nicht im eindimensionalen Sinn einer mehr oder weniger guten Kundenorientierung.
5. Erzeugt man die Skalenwerte der Baumärkte mit einem bestimmten Skalierungsmodell, so ergeben sich aus dem Modell Hinweise für eine systematische Prüfung der Frage, ob eine solche Skala überhaupt existiert bzw. wie stark bestimmte Voraussetzungen für ihre Konstruktion verletzt sind. Ein Beispiel für eine prüfbare Bedingung wurde unter 4. beschrieben. Global lässt sich zudem häufig prüfen, inwieweit sich aus den Skalenwerten die ursprünglichen Daten rekonstruieren lassen. Gelingt dies in hohem Ausmaß, so spricht das dafür, dass die Voraussetzungen für die Skala weitgehend erfüllt sind.

Erfüllt die Skala solche Prüfbedingungen nicht, so könnte man weiter untersuchen, ob es Untergruppen von Heimwerkern gibt, die jeweils eine einheitliche Sichtweise haben und, falls ja, dann für diese Untergruppen jeweils eigene Skalen konstruieren (*multiple Skalierung*). Man kann aber auch fragen, ob Kundenorientierung tatsächlich ein eindimensionales Merkmal darstellt oder ob man nicht eher davon ausgehen muss, dass es sich aus verschiedenen Teilmerkmalen zusammensetzt (z. B. den Öffnungszeiten oder der Freundlichkeit des Personals). In diesem Fall wäre es angemessener, für die Beurteilung der Kundenorientierung mehrdimensionale Skalen zugrunde zu legen.

Vorgehensweisen wie unter 2. und 3. beschrieben wurden vor allem im Kontext der *Klassischen Testtheorie* präzisiert. Das Vorgehen unter 2. dient dabei der Prüfung, ob eine konstruierte Skala das, was sie misst, auch zuverlässig erfasst (unabhängig davon, was dies nun genau ist). Diese wünschenswerte Eigenschaft einer Skala bezeichnet man als ihre *Reliabilität* (genauer gesagt als ihre *Retest-Reliabilität*; andere Formen sind die *Paralleltest-Reliabilität* und die *Interitem-Konsistenz*; vgl. Bühner, 2004). Demgegenüber erlaubt die Strategie unter 3. die Prüfung der *Validität*, also der Frage, wie gut das Merkmal, das man erfassen will, von der Skala gemessen wird (genauer gesagt die *Konstruktvalidität*, für andere Validitätsformen vgl. ebenfalls Bühner, 2004).

Historisch betrachtet wurden vor allem Vorgehensweisen wie unter 4. und 5. beschrieben unter dem Begriff der Skalierung subsumiert. Es existieren verschie-

dene *Skalierungsmodelle*, die jeweils andere Datenerhebungsmethoden erfordern (z. B. Ratings, Rangreihen oder Paarvergleiche; vgl. Borg & Staufenbiel, 2007) und zu Skalen mit unterschiedlichen Eigenschaften (etwa hinsichtlich des Skalenniveaus) führen. Für diese Modelle wurde untersucht, welche Eigenschaften die Daten haben müssen, damit sie in einer Skala abgebildet werden können und wie man in einer *Verlustfunktion* quantifizieren kann, in welchem Ausmaß die Daten gegen die Annahmen verstoßen. Es interessiert also nicht nur, wie man von den Daten zur Skala kommt, sondern auch, ob eine solche Skala für die gegebenen Daten überhaupt existiert bzw. wie gut die Skala ist.

4 Eindimensionale Skalierung

Im Folgenden werden zwei ausgewählte eindimensionale Verfahren dargestellt, die besonders einflussreich in der Skalierung waren. Tabelle 1 stellt diese und einige hier nicht weiter diskutierte Verfahren im Vergleich zentraler Merkmale dar. Alle aufgeführten Modelle werden in Borg und Staufenbiel (2007) ausführlicher abgehandelt.

Tabelle 1: Merkmale einiger eindimensionaler Skalierungsmodelle

Skalierungsmodell	Daten	Itemcharakteristiken	Skala
Skalogramm-analyse (Guttman-Skalierung)	Dominanzdaten von Personen über Objekte	deterministisch, monoton	Gemeinsame Ordinalskala von Personen und Objekten
Thurstone-Skalierung (Law of comparative Judgment, LCJ)	Forced-choice Paarvergleichsdaten der Objekte	probabilistisch, eingipflig	Intervallskala der Objekte
BTL-Skalierung	Forced-choice Paarvergleichsdaten der Objekte	probabilistisch, eingipflig	Intervallskala der Objekte
Unfolding	Präferenzrangreihen der Objekte	deterministisch, eingipflig	Ordered-Metric-Skala der Objekte
Saaty-Skalierung	Verhältnis Paarvergleichsdaten der Objekte	deterministisch, monoton	Verhältnisskala der Objekte
Rasch-Skalierung	Dominanzdaten von Personen über Objekte	probabilistisch, monoton	Gemeinsame Differenzskala von Personen und Objekten