gehen bei der Testentwicklung ausgearbeitet. Sie enthalten i. allg. zu jeder Frage eine Reihe vorgegebener, *gebundener Antworten,* von denen eine als zutreffend zu kennzeichnen ist. Die standardisierten B.en entsprechen dem Vorgehen der Î Psychometric und haben viele Vorteile psychometrischer Verfahren, z. B. können sie Aussagen in bezug auf Normskalen liefern, und ihre Gütekriterien sind überprüfbar, so daß sie sich optimieren lassen. Das einzelne Urteil über eine bestehende Eigenschaft hat in der Regel die Form einer *Schätzung* (rating, engl. Bewertung). Die bekannteste Form der Schätzung ist die Zensierung von Leistungen und Verhaltensweisen, z. B. in Schulen oder Hochschulen, deren Gütekriterien allerdings selten überprüft werden.

In der Psychodiagnostik werden *Schätzskalen* wie Noten-, Einstufungs- oder Rating-Skalen häufig verwendet, insbesondere bei Erlebens- und Verhaltensweisen sowie bei Eigenschaften, die nicht oder schwer durch Tests objektivierbar sind, z. B. bei Schweregraden als Abweichung von einer Norm, bei Ausmaßen von Merkmalen oder bei Häufigkeiten von Verhaltensweisen. Schätzskalen können *unipolar* oder *bipolar* sein, je nachdem, ob ein Nullpunkt, z. B. die Kategorie neutral, indifferent oder mittelmäßig, an einem Pol oder in der Mitte der Skala liegt. Die Anzahl der Kategorien oder der Stufen beträgt mindestens 2, meist 3, kann aber auch 5 oder eine höhere ungerade Zahl sein. Alle oder bestimmte Kategorien der Schätzskala können verbale Bezeichnungen tragen, und der Schätzvorgang kann insbesondere für Laien erleichtert werden durch Angabe von Beispielen, durch „Anker" oder durch graphische Skalen.

Die Schätzskalen sind i. allg. Ordinalskalen (f Skalentypen). Sie werden häufig als metrische Skalen behandelt, indem man ihre Kategorien mit Zahlen bezeichnet, mit denen algebraische Operationen durchgeführt werden können, z. B. die Durchschnittsbildung. Hierbei wird vorausgesetzt, daß Schätzungen eine direkte, eine *absolute Skalierung* darstellen, d. h., daß der Beurteiler wie ein Meßinstrument funktioniert. — Eine andere Möglichkeit der *Metrisierung von Schätzskalen* bildet die Normalisierung; häufig sind die Ergebnisse der direkten Skalierung und der Normalisierung untereinander konform, d. h., sie stehen in linearer Beziehung.

Das Hauptproblem aller Schätzurteile liegt in der Subjektivität der Beurteilung, die sich in verschiedenen Fehlerquellen bzw. Verzerrungstendenzen äußert, z. B. kann der Beurteilermaßstab an bestimmte Populationen gebunden sein, etwa an Hilfsschüler, Oberschüler oder an verschiedene Altersstufen; eine Übertragung des Urteils von einer Eigenschaft auf eine andere bezeichnet man als *Hof-Effekt; Milde-* oder *Strenge-Effekt* kennzeichnen persönliche Urteilstendenzen bestimmter Beurteiler; eine Abhängigkeit der Schätzung von bestimmten Annahmen über Zusammenhänge zwischen Eigenschaften ergibt den *Logik-Effekt,* die Bevorzugung mittlerer Schätzungen den *Extremscheu-Effekt,* Urteile, die nach persönlichen Wünschen oder sozialen Erwartungen verzerrt sind, den *Effekt der Normanpassung oder der Erwünschtheit.* Die *Objektivität* (f Reliabilität) von Schätzungen läßt sich überprüfen durch Wiederholung der Schätzung, d. h. durch *Intra-Urteiler-Invarianz,* oder durch mehrere Beurteiler, d. h. durch *Inter-Urteiler-Invarianz.* Durch diese Verwendung mehrerer Schätzungen wie auch durch Training der Beurteiler läßt sich die Objektivität bzw. die intersubjektive Übereinstimmung erhöhen. Reliabilitäts- und Validitätskoeffizienten von Einzel-Beurteilern liegen zumeist unter denen von Tests.

Eine weitere Methode der B. ist die *Rangordnung,* d. h. die Ordnung einer Stichprobe gemäß der Reihenfolge der Vpn. in bezug auf die Ausprägung einer bestimmten Eigenschaft. Die Rang-Skala ist ebenfalls eine Ordinalskala, die sich unter der Voraussetzung der Normalverteilung in eine Standardskala transformieren läßt. In der Psychometrie sind die Anwendungsmöglichkeiten der Rang-Skala begrenzt, da sie nur bei kleinen Stichproben praktikabel sind und ihr Ergebnis an die betreffende Stichprobe gebunden ist. Sie wird häufig verwendet, wenn psychologische Laien relativ kleine, ihnen gut bekannte Gruppen von Vpn. zu beurteilen haben.

Ähnliches gilt für die dritte B.s-Methode, den *Paarvergleich.* Bei diesem für sehr verschiedenartige psychologische Fragestellungen — insbesondere der j Psychophysik — geeigneten Verfahren werden sukzessiv Vergleiche in bezug auf alle aus einer Gruppe von Vpn. oder Objekten zu bildenden Paare angestellt. Der einzelne Paarvergleich ist in der Regel einfach, durch die bei n Vpn. erforderlichen $[ff) = n(n — X)I2$ Vergleiche aber zumeist mühsam, z. B. 45 Vergleiche für $n = 10$, 190 für n = 20, 435 für $n = 30$ usw. Zwischen den drei genannten Beurteilungsmethoden besteht insofern ein formaler Zusammenhang, als die Schätzungen einer Eigenschaft für eine Gruppe von n Vpn. eine Rangordnung und — über die Dominanzmatrix — $\binom{n}{2}$ Paar vergleiche impliziert. Die umgekehrte Beziehung gilt nicht; jedoch lassen sich mittels aller drei Verfahren Standard-, d. h. metrische Skalen herstellen.

Bewährung: 1. das Eintreffen einer Vorhersage, die auf Grund von diagnostischen Feststellungen getroffen worden ist. Die Größe dieser B. wird als die prognostische Validität der diagnostischen Methode bezeichnet und als Korrelation zwischen Testergebnis und B.skriterium angegeben. — 2. die Stabilität eines durch pädagogische oder therapeutische Einwirkungen zustande gekommenen Effekts.