Validität (= Gültigkeit): Genereller Terminus, der die Richtigkeit einer Messung bezeichnet (damit ist gemeint, daß ein Test tatsächlich das mißt, was er vorgibt zu messen).

Inhaltliche Validität: Das Ausmaß, in dem Testaufgaben den inhaltlichen Bereich abdecken, den der Test vorgibt zu messen. Bei Leistungstests z.B. ist die inhaltliche Gültigkeit von entscheidender Bedeutung. Ein Beispiel für einen Test mit geringer inhaltlicher Gültigkeit wäre ein Mathematiktest für Zehntkläßler, der (zu schwierige) Berechnungen aus der Höheren Mathematik enthält, oder viele (für diese Stufe zu einfache) Multiplikationsaufgaben von zweistelligen Zahlen, oder der komplexe Textaufgaben beinhaltet, die so kompliziert und verschachtelt formuliert sind, daß ein Teil der Testanden sie allein schon deshalb nicht lösen kann, weil er die Aufgabentexte gar nicht versteht.

Empirische Validität: Hiermit werden alle Validationsverfahren bezeichnet, die die Testwerte einer Stichprobe mit anderen unabhängigen Testwerten (= Kriteriumswerten) der gleichen Stichprobe vergleichen. So kann z.B. die Notengebung eines Lehrers validiert werden, indem man seine Noten mit dem Durchschnitt aller anderen Noten korreliert (=vergleicht). Wenn beispielsweise viele Schüler, die im Vergleich zu den übrigen Schülern der Klasse gute Noten in den Sprachfächern Deutsch und Englisch haben, schlechte Noten im anderen Sprachfach Französisch bekommen, dann ist die Notengebung des Französischlehrers wahrscheinlich empirisch nicht valide. Läßt sich die gleiche Diskrepanz auch in den anderen Klassen des gleichen Lehrers feststellen, müßte er eigentlich in sich gehen.

Konkurrierende Validität: Das Ausmaß, in dem ein Test hoch korreliert (= übereinstimmt) mit einem anderen allgemein anerkannten Meßinstrument (z.B. Expertenurteil oder Test) für das untersuchte theoretische Konstrukt (= Eigenschaft, Fähigkeit, Wissen …). So hat z.B. die neue (und damit ökonomische) Kurzform eines Intelligenztests konkurrierende Validität, wenn die Ergebnisse des Kurztests sehr ähnlich sind im Vergleich zu dem Abschneiden in einem umfangreicheren anerkannten Meßverfahren wie z.B. dem Stanford-Binet oder Hamburger Intelligenztest. Daß das Konzept der (konkurrierenden) Validität graduell zu verstehen ist und stochastisch (=wahrscheinlichkeitstheoretisch) berechnet wird, zeigt sich daran, daß in der empirischen Wirklichkeit zwei Testergebnisse ganz selten mal identisch sind (hundertprozentig übereinstimmen). Erwartbar ist in der Regel nur eine mehr oder weniger ausgeprägte Ähnlichkeit, d.h. relative Übereinstimmung.

Konstrukt-Validität: Bezeichnet das Ausmaß, in dem die gemessenen Testergebnisse mit theoretischen Erwartungen übereinstimmen.
Will man z.B. einen neuen Test für Konservativismus validieren, so kann man die Ergebnisse dieses Tests korrelieren mit den Ergebnissen eines allgemein anerkannten Tests für Autoritätsgläubigkeit. Gemäß der theoretischen Annahme, daß autoritätsgläubige Menschen meist auch politisch konservativ sind, müßten die Ergebnisse in beiden Tests stark übereinstimmen.

Externe Validität: Bezeichnet das Ausmaß, in dem man die Ergebnisse eines Experiments/Tests verallgemeinern kann. Die externe Validität ist eine Funktion des Grades, mit dem in einer Untersuchung die Personen, die Rahmenbedingungen und die Verfahren übereinstimmen mit einem "natürlichen" oder "realen" Phänomen, das experimentell modellhaft simuliert wird.

Interne Validität:(1) In der *Testtheorie* das Ausmaß, in dem ein Test das mißt, was er beabsichtigt zu messen. (2) In der *empirischen Methodenlehre* das Ausmaß, in dem man eine Ursache-Wirkung-Beziehung nachweisen kann zwischen einer unabhängigen Variablen (– die man im Experiment kontrolliert verändert) und einer davon abhängigen Variablen (– deren Veränderung man mißt). So kann ich z.B. bei einer Nacherzählung die Korrelationen berechnen zwischen dem (mir bekannten, weil vorher gemessenen) Abschneiden der Schüler bei einem Vokabeltest (Beeinflussung der Nacherzählung durch den verfügbaren Wortschatz des Schülers) und bei einem Grammatiktest (der ja die Fähigkeit des Schülers zu einem morphologisch und syntaktisch korrekten Sprachgebrauch ausdrückt). Es gibt auch fortgeschrittene Verfahren, wie man die relative Stärke des Einflusses von Vokabelkenntnissen und Grammatikwissen auf die Güte der Nacherzählung auseinanderrechnen kann. Diese statistischen "Trennverfahren" sind notwendig, da in der experimentellen Alltagspraxis immer wieder Variablen miteinander konfundiert (= vermischt, verwechselt) werden. Im oben geschilderten Fall wäre es z.B. gut möglich, daß ein gutes Gedächtnis für die Übereinstimmung beim Vokabeltest, der Grammatikarbeit und der Nacherzählung gleichzeitig ursächlich verantwortlich ist.

Vorhersage-Validität: Bezeichnet das Ausmaß, wie ein Testergebnis zukünftige Ergebnisse vorhersagt. So kann man beispielsweise bei Studierenden der Anglistik oder der Romanistik das Abschneiden bei einem fremdsprachlichen Einstufungstest mit der Note in der Übersetzungsklausur beim Staatsexamen vergleichen.

Reliabilität (= Zuverlässigkeit): Ein allgemeiner Terminus, der die Konsistenz (Übereinstimmung) zwischen Meßergebnissen bezeichnet, die bei wiederholten Beobachtungen der gleichen Versuchspersonen unter gleichen Umständen festgestellt werden. Eine hohe Reliabilität erhöht die Verläßlichkeit (der Ergebnisse) eines Tests, weil die Ergebnisse in geringerem Maße von zufälliger Variabilität (Veränderlichkeit, Streuung) beeinflußt werden. Die Maßzahl für die Reliabilität ist ein Koeffizient, berechnet über Korrelation oder Stärke der Assoziation (Verknüpfung, räumliche und/oder zeitliche Nähe).

Parallelform-Reliabilität: Verfahren zur Erfassung der Reliabilität einer Messung. Eine ähnliche oder alternative (= parallele) Form eines zuvor implementierten Tests mit den gleichen Aufgabentypen wird nach einer angemessenen Zeit implementiert. Die Ergebnisse beider Testformen werden korreliert, um die Reliabilität zu berechnen.

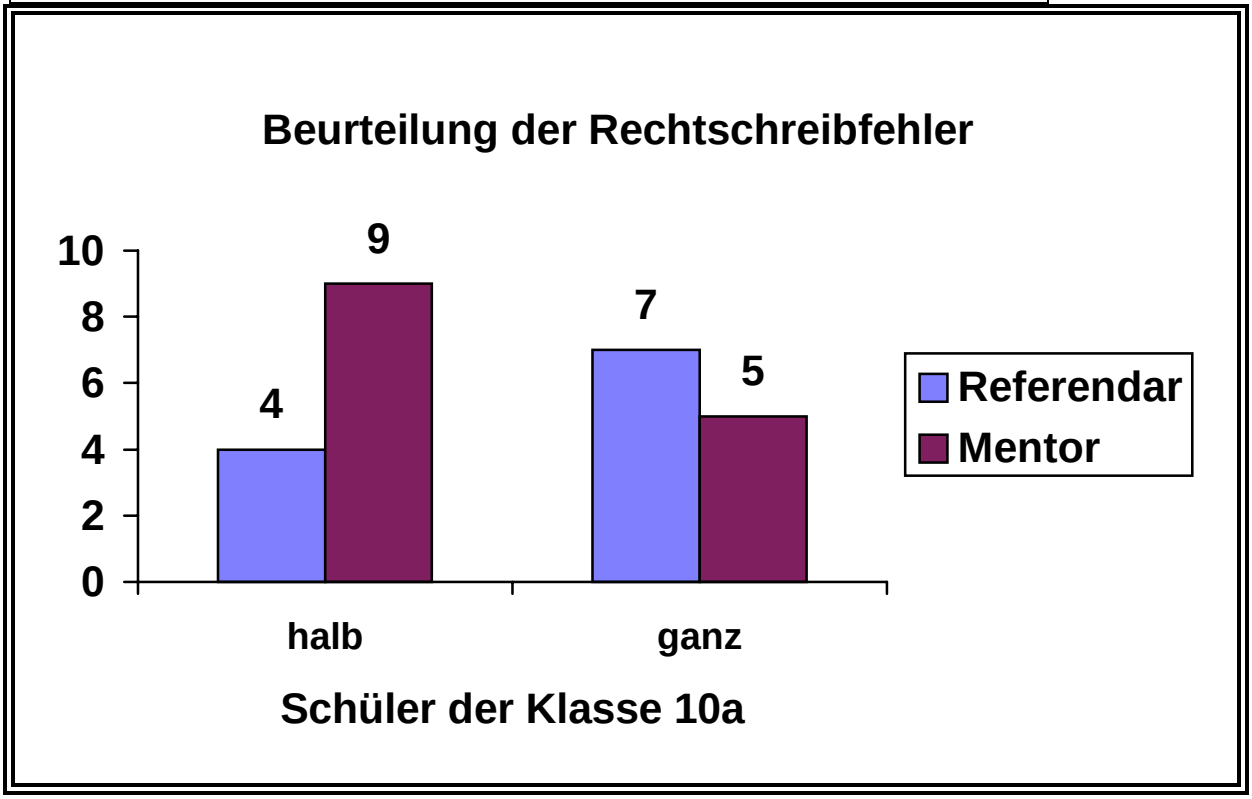
Halbierungsreliabilität: Ist ein Reliabilitätsmaß für die interne Konsistenz eines Tests. Man erhält es, indem man das Ergebnis bei der einen Hälfte der Items mit dem Abschneiden bei der anderen Hälfte vergleicht. Das am häufigsten verwendete Verfahren ist die Berechnung der Korrelation zwischen den Ergebnissen bei den Testaufgaben mit gerader Laufnummer vs. mit ungerader Laufnummer.

Wiederholungs-Reliabilität: Sammelbezeichnung für alle Verfahren, die die Reliabilität eines Tests dadurch messen, daß sie die Ergebnisse der gleichen Testandengruppe bei der wiederholten (i.d.R. nochmaligen) Implementation desselben Tests miteinander korrelieren. Das Verfahren kann freilich nur angewendet werden, wenn während der jeweils voraufgegangenen Implementation keine Lern- oder Gedächtniseffekte aufgetreten sind.

Interreliabilität (*syn*. Übereinstimmung zwischen Auswertern): Das Ausmaß, in dem zwei oder mehrere Auswerter über die charakteristischen Ausprägungen von beobachteten Eigenschaften übereinstimmen.
Die Maßzahl für die Interreliabilität ist der ***Phi Koeffizient (ϕ)*** [eine synonyme Bezeichnung ist: ***Vierfache Punktkorrelation (r_b)***].

Fehler in einem Diktat in der 10a	Bewertung als halber Fehler	Bewertung als ganzer Fehler	Σ
Referendar	a) 4	b) 7	11

Mentor	c) 9	d) 5	14
Σ	13	12	25



Berechnung von Phi nach der Formel:

$$\phi = ad - bc / \sqrt{(a+b)\,(c+d)\,(a+c)\,(b+d)} = 20\text{-}63 / \sqrt{(11)\,(14)\,(13)\,(12)}$$

$$= -43 / \sqrt{24024} = -43 / 155 = -0{,}28$$

Das negative Vorzeichen zeigt, daß die beiden Lehrer die Rechtschreibfehler gegenläufig beurteilen. Allerdings ist diese Gegenläufigkeit nur schwach ausgeprägt (die Lehrer verwenden offensichtlich unterschiedliche Kriterien für ihre Beurteilung). Wie bei einem Korrelationskoeffizienten läßt sich das Quadrat von Phi interpretieren als prozentuale Übereinstimmung ($-0{,}28^2 = 8\%$). M.a.W. Die Interreliabilität bei der Beurteilung der Rechtschreibfehler als ganzer Fehler (z.B. als Indiz für fehlende Wortschatz- oder Grammatikkenntnis) oder als halber Fehler (halt nur falsch geschrieben, vielleicht sogar nur ein Flüchtigkeitsfehler) in der Klasse 10a durch die beiden Lehrer ist so gering, daß sie für praktische Konsequenzen kaum verwendet werden kann. Weder im positiven noch im negativen Sinne.

Freilich ist nicht zu unterschätzen, daß ein solches Rechenergebnis eine Diskussion über die Bewertung von Rechtschreibfehlern nicht nur notwendig macht, sondern darüber hinaus auch auf eine rationale und transparente Basis stellt. Die ja nachgewiesene Notwendigkeit einer Einigung läßt sich nicht mehr nur über vermutete oder durch Erfahrung gewonnene Autorität erzielen (die allerdings viel zu beliebte pädagogische Einbahnstraße), die Einigung läßt sich

nach einer objektiven quantitativen Analyse nur mehr über die explizite Begründung von akzeptablen Auswertungs- und Bewertungskriterien herbeiführen.