25.3 Angaben zur Validität

Die Qualität eines Diagnoseverfahrens hängt besonders von seiner Validität ab. Ein Verfahren, dass zwar unter standardisierten Bedingungen sehr zuverlässig Daten erhebt, weist sich nicht allein schon dadurch als ein brauchbares Verfahren aus. Erst wenn es nachweislich tatsächliche Merkmale des interessierenden Verhaltens erfasst, für dessen Beobachtung es konzipiert ist, kann es für dieses Verhalten als gültiges Untersuchungsinstrumentarium betrachtet werden. Dabei ist zwischen einer internen und einer externen Validität zu unterscheiden.

Interne Validität

Hinsichtlich der internen Validität besteht ein enger Zusammenhang zwischen der inhaltlichen oder logisch erklärbaren Validität und der sogenannten Konstruktvalidität, die sich nicht etwa auf die Konstruktion des Verfahrens bezieht, dessen Materialien und Untersuchungsanordnungen auf Objektivität und Reliabilität zu überprüfen sind, sondern auf ein theoretisches Konstrukt. Im Fall der Sprachkompetenz sind aus der Spracherwerbsforschung abzuleitende Erkenntnisse über Aspekte des komplexen Sprachverhaltens relevant, die bereits im Teil I ausführlich dargestellt wurden. Da bei diesem Verfahren zur Erfassung einer bilingualen Sprachkompetenz unter Beachtung altersgemäß zu erwartender Leistungen die Grundfähigkeiten der Sprache wie Hören, Sprechen und Lesen in beiden Sprachen sowie zusätzlich das Schreiben auch in der starken Sprache beobachtet wurden, dürfte eine ausreichende Konstruktvalidität vorliegen.

Inwieweit die dargebotenen Untersuchungsanordnungen geeignet waren, diese Sprachfähigkeiten auch auf allen beim Sprachverhalten zu unterscheidenden Ebenen (Phonematik, Lexematik, Morphematik usw.) zu evozieren, ist indessen nicht zu belegen. Der Schwerpunkt lag bei Beobachtungen zum mündlichen Sprachgebrauch über integrierte Aufgabenstellungen, um die Sprachäußerungen möglichst unter Bedingungen zu erfassen, die einer natürlichen Kommunikation ähneln. Dennoch konnte die Kommunikationsfähigkeit so kaum in realen Situationen überprüft werden. Beim Hören und Sprechen wurden die sprachlichen Fähigkeiten erst bei der Auswertung einzeln auf der Laut-, Wort- oder Satzebene betrachtet. Die integrierte Datenerhebung erscheint der Beobachtung von normalerweise kontextabhängigem Sprachverhalten aber angemessener als isoliertes Abfragen einzelner Bereiche. Dagegen tragen die in Sprachtests beliebten Darbietungsformen nicht unbedingt zur Erhöhung der inhaltlichen Validität, sondern oftmals zur Verzerrung bei. So hängt bei vielen *Paper und Pencil*-Tests, bei denen grammatikalische Kenntnisse durch

schriftliche Eintragungen in Lücken überprüft werden, die Performanz auch von der Schreibfertigkeit ab und bei *multiple choice*-Aufgaben ist stets fraglich, wie viel richtige Lösungen durch Raten auftreten.

Die interne Validität eines Verfahrens ist im Allgemeinen durch statistische Verfahren nur unvollkommen quantifizierbar. Die Ergebnisse der Untersuchung könnten etwa mit Daten verglichen werden, die bei der Anwendung eines zweiten Sprachdiagnose-verfahrens gewonnen wurden, was mangels eines anderen Verfahrens zur Erfassung deutsch/ italienischer Sprachfähigkeiten im Primarbereich jedoch entfällt. Ein solcher Vergleich würde auch nur eine vage Einschätzung zulassen, da eine solche Validierung wiederum von der Validität des anderen Verfahrens abhängig wäre. Auch die Faktorenanalyse, mit der konstruktnahe und konstruktferne Teile identifizierbar sind, kann nur einen Aspekt der Konstruktvalidität aufklären. Daher wird über die inhaltliche Validität eines Verfahrens üblicherweise durch ein Rating von Experten befunden. Insofern obliegt bei diesem noch nicht veröffentlichten Bericht zunächst den Gutachtern die Entscheidung darüber, ob sie die im Verfahren benutzten Aufgaben und Bewertungskriterien als eine hinreichend repräsentative Auswahl für das Konstrukt Sprachverhalten einschätzen.

Lienert beschreibt in seinem Standardwerk eine Methode,[256] mit dem verfahrensintern ein Validitätskennwert ermittelt werden kann, der zwar nicht die Höhe der Validität anzugeben vermag, aber eine gesicherte Aussage über die Validität des Verfahrens zulässt, wenn der errechnete *t* - Wert einer Irrtumswahrscheinlichkeit von weniger als *p* = 1% entspricht. Bei dieser Extremgruppenmethode werden die Ergebnisse von zwei Probandengruppen miteinander verglichen, die das beobachtete Merkmal in hohem bzw. nur in niedrigem Grade aufweisen. In der vorliegenden Untersuchung wurde diese Probe bezüglich der italienischen Sprache mit 7 Kindern aus der italienischen (20,5 Punkte und höher) und 4 Kindern aus der deutschen Sprachgruppe mit einem Ergebnis unter 9 Punkten durchgeführt. In Bezug auf die deutsche Sprache wurden 5 Kinder der deutschen Sprachgruppe mit Ergebnissen über 20 Punkten (jeweils Bezug 22 ohne Schreibfertigkeit wegen des Vergleichs mit Partnersprache) ausgewählt und mit den beiden Schülern der italienischen Sprachgruppe verglichen, die in ihrer Zweitsprache Deutsch gegen Ende des 2. Schuljahres erst zirka 13 Punkte erreichten (Berechnung siehe Tabellenblatt 22). Demnach ergibt sich bei beiden Sprachen für das an der SESB eingesetzte Verfahren jeweils ein Validitätskennwert mit $p < 1$ %,

[256] Lienert 1969, a.a.O., S.280ff – Meines Erachtens ist bei einer Validierung über die Extremgruppenmethode jedoch eine gewisse Skepsis angebracht. Angesichts der bei der Berechnung überaus hoch ausfallenden Signifikanzstufe habe ich die Methode auch auf fiktive Daten angewandt, wobei festzustellen war, dass sich nur beim gleichzeitigen Auftreten hoher Standardabweichungen und einer geringen Differenz der Mittelwerte keine Signifikanz ergab.

womit nach Lienert das Verfahren als valide anzusehen ist. Da mit dieser Methode aber letztlich nur im Sinne des Reliabilitätskonzepts ermittelt wird, ob das Verfahren zwischen starken und schwachen Sprachfähigkeiten unterscheiden kann, sollen die Angaben zur Validität noch durch die Prüfung der externen Validität ergänzt werden.

Externe Validität

Die externe Validität wird auch kriterienbezogene Validität genannt, weil sie anhand eines Außenkriteriums bestimmt wird. Dabei wird überprüft, inwieweit die Ergebnisse verallgemeinerbar sind bzw. ob sich die Diagnose in der Praxis bestätigt. Da es sich bei den Probanden dieser Untersuchung um an einem Schulversuch teilnehmende Kinder handelt, kommt eine Validierung anhand des Lehrerurteils in Frage, das allerdings kein sehr objektives Außenkriterium darstellt, denn bekanntlich beurteilen Lehrer die Leistungen ihrer Schüler mitunter auch nach subjektiven Kriterien. Bei der Validierung über ein Außenkriterium hängt das Ergebnis nicht nur von der Reliabilität des Verfahrens, sondern in starkem Maße auch von der Reliabilität des Kriteriums, also hier des Lehrerurteils ab.

Bei dieser Untersuchung wurden alle beteiligten Lehrer um die Einschätzung der Sprachkompetenz ihrer Schüler in der starken Sprache und der Partnersprache (SESB) hinsichtlich des Gesamteindrucks gebeten. Die Beurteilung zu einzelnen Sprachbereichen wurde nicht erfragt. Anhand der folgenden Kriterien waren absichtlich keine Ziffern, die leicht die Assoziation zu Schulnoten hervorgerufen hätten, sondern Buchstaben zuzuordnen.[257] Da auch die Beurteilung in Zwischenschritten möglich war, standen neun Kategorien zur Verfügung.

mindestens altersgemäße oder überdurchschnittliche Sprachkenntnisse	A
fast altersgemäße Sprachkenntnisse mit nur geringen Schwächen	B
Kommunikation gelingt. Äußerungen weisen aber noch erhebliche Schwächen auf.	C
Bemerkbare Fortschritte gehen über ein Anfängerstadium hinaus.	D
noch im Anfängerstadium, Fortschritte sind kaum bemerkbar	E

[257] Auch so ist die bei einigen Lehrern zu beobachtende Zurückhaltung hinsichtlich „schlechter" Beurteilungen nicht auszuschließen, aber eine Orientierung an Schulnoten wäre noch stärker subjektiven Urteilen unterworfen. Z. B. hätte die Kategorie D nach der Reihenfolge einer 4 entsprochen, die als Beurteilung für *ausreichende* Leistungen gilt. Aber in Bezug auf altersgemäße Sprachnormen wäre dieses Stadium noch nicht als *ausreichend*, in Hinsicht auf einen Zweitspracherwerb seit Schulbeginn jedoch schon als *gut* zu bezeichnen. – Bei der statistischen Auswertung wurde dem Niveau A abweichend von der in Deutschland üblichen Benotung der höchste Punktwert 5 der mehrstufigen Skala zugeordnet.

Einschätzungen durch einen einzigen Beurteiler weisen relativ oft große Zufallsfehler auf, da sie von seinem Einschätzungsvermögen und der eher wohlwollenden, strengen oder durchschnittlichen Beuteilungstendenz abhängen. Auch bei den um ihr Urteil gebetenen Lehrern weicht das Einschätzungsverhalten stark voneinander ab. Bei den anhand der Korrelation zwischen Lehrerurteil und Untersuchungsergebnis ermittelten Validitätskoeffizienten von r_{tc} = 0.65 bei der starken Sprache und r_{tc} = 0.84 bei der Partnersprache wären daher Korrekturen angebracht, welche die Unreliabilität des Außenkriteriums berücksichtigen. Statt der ziemlich umständlichen Korrekturberechnungen wird hier jedoch ein Kommentar bevorzugt, der das Einschätzungsverhalten analysiert und deutlich nahe legt, die Validitätskoeffizienten wegen offensichtlich eingeschränkter Reliabilität des Lehrerurteils nur als relative Schätzung aufzufassen. Deshalb werden in die tabellarische Übersicht der Korrelationen zwischen dem Untersuchungsergebnis und den Lehrerurteilen auch Interkorrelationen aufgenommen, die weitere Informationen bieten. Anhand eines Vergleichs der verfahrensinternen Interkorrelationen mit denen zwischen dem jeweiligen Lehrerurteil und einzelnen Untersuchungsbereichen kann nachvollzogen werden, inwieweit die Lehrer bei ihrer allgemeinen Beurteilung des Sprachverhaltens eine ähnliche Vielfalt der Sprachkompetenz bedacht haben und welche Bereiche in ihrer Urteilsfindung hinsichtlich der Gewichtung abweichen.[258]

Bei der starken Sprache korrelieren die Lehrerurteile besonders hoch bei beiden Kontrollgruppen und der italienischen Sprachgruppe der Klasse y, wobei aber bei der Lehrerin der italienischen Kontrollgruppe die Beurteilung nur in 4 von 9 Bereichen auch dem Varianzanteil des Untersuchungsverfahrens weitgehend entspricht (weitgehende Entsprechung = Abweichung um höchstens 0,15), was bei den beiden andern Lehrern immerhin in 6 Bereichen der Fall ist. Abgesehen von dem sehr unterschiedlichen Einbezug der Lesefertigkeit in das Urteil, zeigen die Abweichungen einige Gemeinsamkeiten auf, z.B. achten anscheinend alle drei Lehrer beim mündlichen Sprachgebrauch stärker auf phonetisch-prosodische Merkmale der Sprache und halten die Schreibfertigkeit für wichtiger, als das Verfahren im 2. Schuljahr vorsah. Der Lehrer der deutschen Kontrollgruppe legt überdies größeren Wert auf die Lesefertigkeit, während die italienische Lehrerin der Klasse y mehr auf morphosyntaktisch korrekte Sprachäußerungen achtet. Beim Lehrerurteil der italienischen Kontrollgruppe werden die Lesefertigkeit und das kommunikative Sprachverhalten kaum berücksichtigt.

[258] Da bei den folgenden Ausführungen zum Kriterium des Lehrerurteils unter Beibehaltung der originären Klassenbezeichnungen zwangsläufig Rückschlüsse auf die Personen möglich wären, werden die Klassen der SESB in diesem Abschnitt durch imaginäre Bezeichnungen unterschieden. Die betroffenen Lehrer der Kontrollgruppen mögen die Indiskretion verzeihen.

Interkorrelationen des Lehrerurteils (starke Sprache) mit Untersuchungsbereichen und dem Gesamtergebnis

	starke Sprache Deutsch					starke Sprache Italienisch					D + It	Interkorrel.
	SESB Klasse x	SESB Klasse y	SESB Kl. x+y	Kontroll-gruppe	alle D	SESB Klasse x	SESB Klasse y	SESB Kl. x+y	Kontroll-gruppe	alle It.	N = 55	Verfahren/ Gesamterg.
Gesamtergebnis	0,50	0,27	0,51	0,87	0,78	0,82	0,82	0,47	0,90	0,55	0,65	(1,00)
Phonem./Pros. G	0,61	0,00	0,54	0,97	0,87	0,42	0,93	0,19	0,89	0,41	0,62	0,62
Textverständnis	0,16	0,00	0,11	0,54	0,56	0,83	0,77	0,72	0,27	0,63	0,62	0,62
Hörverständnis G	-0,20	0,00	- 0,13	0,73	0,65	0,67	0,77	0,65	0,41	0,58	0,62	0,62
Schreibfertigkeit	0,21	0,00	0,30	0,77	0,62	0,03	0,75	0,30	0,59	0,39	0,47	0,47
Lesefertigkeit	0,38	0,00	0,44	0,94	0,86	0,14	0,47	-0,03	0,53	0,12	0,49	0,49
Wortschatz G	0,33	-0,65	0,38	0,53	0,43	0,69	0,46	0,44	0,33	0,40	0,46	0,46
Morphosyntax G	0,33	-0,37	0,40	0,47	0,47	0,71	0,93	0,50	0,65	0,53	0,50	0,50
Konzepte	0,89	0,78	0,76	0,34	0,29	0,61	0,45	0,44	0,11	0,33	0,36	0,36
Komm. Sprachv.	0,43	0,58	0,27	0,44	0,22	0,39	0,60	0,29	0,21	0,27	0,21	0,66

Dafür traf die deutsche Lehrerin der Klasse y ihre Entscheidung fast ausschließlich aufgrund ihres Eindrucks vom kommunikativen Verhalten, d.h. ohne Beachtung anderer Sprachfähigkeiten, obwohl besonders im Bereich der starken Sprache um die Beurteilung umfassender Sprachkenntnisse gebeten worden war. Die Korrelation ihres Urteils mit dem Gesamtergebnis fällt dementsprechend besonders niedrig aus (0,27), die Korrelationen zu sprachlichen Teilbereichen liegen bei Null oder nehmen negative Werte an. Nur der Bereich Begriffsbildung (Konzepte), der aber eher die kognitive als die sprachliche Entwicklung erfasst, korreliert noch besonders hoch mit ihrem Urteil. Da sprachspezifische Bereiche offensichtlich überhaupt nicht berücksichtigt wurden, spricht die niedrige Korrelation in diesem Fall auch weniger gegen die Validität des Verfahrens als gegen die Reliabilität des Lehrerurteils als Außenkriterium. Weil diese Lehrerin Sprachniveaus im starken Bereich anscheinend nur schwer unterscheiden kann, fielen ihre Beurteilungen entsprechend undifferenziert aus. Die hohe Korrelation von 0,81 im Bereich der Partnersprache ergibt sich denn auch nur zufällig aus dem Umstand, dass die deutschen Sprachfähigkeiten der von ihr zu beurteilenden Gruppe auch nach dem Verfahren als ziemlich gleich stark resultieren und eine Differenzierung hier also nicht erforderlich war. Ansonsten berücksichtigte sie anscheinend – wie schon bei der starken Sprache – kaum Aspekte des mündlichen Sprachgebrauchs, aber sehr viel stärker die Aussprache. Absurderweise korreliert ihre Einschätzung bei der Partnersprache auch stärker mit der Lesefertigkeit als bei der starken Sprache, obwohl die verfahrensinterne Korrelation bei der Partnersprache einen geringeren Anteil der Varianzaufklärung durch die Lesefertigkeit nahe legt.

Bei den beiden anderen Lehrern, deren Beurteilung bei der starken Sprache weder besonders schwach noch besonders hoch (0,50 + 0,62) mit dem Gesamtergebnis korreliert, fällt auf, dass sie sich in der Beurteilung der Sprachfähigkeit vermutlich auch von ihrem Eindruck der Lernfähigkeit beeinflussen lassen. Dafür sprechen die hohen Korrelationen mit dem Konzepte-Fragenkatalog. Ansonsten legen sie beide im 2. Schuljahr anscheinend noch wenig Wert auf die Kulturtechniken Lesen und Schreiben. Übrigens hat die zaghafte Beurteilung der italienischen Lehrerin der Klasse x größere Auswirkungen auf die Übereinstimmungsquote als die mangelnde Differenzierung der deutschen Kollegin der Klasse y. Anscheinend hatte die italienische Lehrerin Skrupel, allzu schlechte Urteile zu fällen, wodurch sich statistisch bei Zusammenfassung der italienischen Gruppen teilweise niedrigere Korrelationen ergeben als bei Einzelbetrachtung der Gruppen. Innerhalb ihrer Lerngruppe haben zwar beide italienische Lehrerinnen der SESB Differenzierungen vorgenommen, die der Graduierung der Untersuchungsergebnisse entsprechen, doch weichen die gewählten Niveauabstufungen stark voneinander ab. Da die italienische Lehrerin der Klasse x auch dem schwächsten Schüler (Untersuchungsergebnis 10,88 bei Bezug 25) ihrer Gruppe noch eine ausreichende Sprachkompetenz zuerkennt, führt diese Fehleinschätzung unweigerlich zu einer Minderung der Korrelation, sobald auch die stimmigeren Beurteilungen der Kollegin einbezogen werden, denn dann ändert sich das Übereinstimmungsverhältnis zwischen Lehrerurteil und Untersuchungsergebnis erheblich.

Interkorrelationen des Lehrerurteils (Partnersprache) mit Untersuchungsbereichen und dem Gesamtergebnis

	Partnersprache Deutsch			Partnersprache Italienisch			D + It	Interkorrel.
	SESB Klasse x	SESB Klasse y	SESB Kl. x+y	SESB Klasse x	SESB Klasse y	SESB Kl. x+y	N = 36	Verfahren/ Gesamterg.
Gesamtergebnis	0,70	0,81	0,71	0,65	0,77	0,74	0,84	(1,00)
Phonem./Pros. G	0,66	0,72	0,68	0,20	0,55	0,47	0,73	0,89
Wortschatz Bild	0,61	0,29	0,56	0,37	0,80	0,66	0,75	0,84
Morphosyntax Bild	0,89	0,29	0,75	0,52	0,82	0,68	0,80	0,90
Mündl. Sprachg. G	0,68	0,28	0,90	0,75	0,87	0,81	0,83	0,97
Hörverständnis G	0,52	0,46	0,47	0,77	0,73	0,77	0,75	0,90
Phon./Pros. Lesen	0,26	0,72	0,38	0,15	0,60	0,47	0,41	0,57
Lesefertigkeit	0,12	0,65	0,31	-0,14	0,16	-0,01	0,14	0,34
Konzepte	0,40	0,54	0,49	-0,11	0,30	-0,16	0,04	0,04
Komm. Sprachv.	0,50	0,61	0,53	0,50	0,00	0,45	0,66	0,72

Auch bei der Beurteilung der Partnersprache, in der graduelle Unterschiede wegen der meistens heterogeneren Lernvoraussetzungen leichter zu erkennen waren, ist ein sehr unterschiedliches Einschätzungsverhalten zu beobachten. Dabei verwendeten die italienischen Lehrerinnen jedoch ziemlich konsistent ihren subjektiven Maßstab, auf

dessen Grundlage sie auch die Kompetenz in der starken Sprache beurteilt haben. Der Lehrerin der Klasse x waren z. B. auch in der Partnersprache die Aussprache und die Lesefertigkeit weitaus weniger wichtig als das Hörverständnis, der mündliche Sprachgebrauch und das kommunikative Sprachverhalten. Bei der Beurteilung der Klasse y wurden wiederum viele der auch vom Verfahren vorgesehenen Aspekte in ähnlicher Gewichtung einbezogen, d.h. die Lesefertigkeit kaum, desto stärker aber das Hörverständnis, der Wortschatz sowie grammatikalische und phonetisch-prosodische Merkmale der Sprache. Allein das Verhältnis des kommunikativen Aspekts entspricht nicht dem verfahrensinternen Varianzanteil. Das mag daran liegen, dass die Lehrerin der Klasse y in besonderem Maße auf die Korrektheit der Sprachäußerungen achtet und sich wenig von einer „Vielplauderei“ beeindrucken lässt.

Beide italienische Lehrerinnen haben sich übrigens im Gegensatz zu ihren deutschen Kolleginnen bei der Beurteilung der Sprachkompetenz in der Partnersprache nicht von ihrem Eindruck der allgemeinen Lernfähigkeit oder des Allgemeinwissens der Schüler blenden lassen, sondern ihre Entscheidungen auf sprachimmanente Beobachtungen gestützt. Das ist den niedrigen bzw. negativen Korrelationen des Bereichs Konzepte mit ihrem Urteil zu entnehmen, der hier sozusagen als Kontrollvariable aufgeführt ist. Bei der Klasse x hat die deutsche Lehrerin aber außerdem auch sprachspezifische Aspekte berücksichtigt, während die Beurteilungen bei der Klasse y in Bezug auf die deutsche Sprache, wie schon erwähnt, durchweg als wenig sprachbezogen und undifferenziert auffallen.

Aufgrund der dargestellten Unterschiede hinsichtlich des Einschätzungsverhaltens ist davon auszugehen, dass die Reliabilität des Außenkriteriums Lehrerurteil bei dieser Untersuchung äußerst fragwürdig ist. Angesichts des recht unzuverlässigen Kriteriums ist allerdings auch die externe Validierung dieses Verfahrens stark beeinträchtigt und daher nur mit Vorbehalt zu betrachten. Da die hier ermittelten Validitätskoeffizienten von $r_{tc} = 0.65$ bei der starken Sprache und $r_{tc} = 0.84$ bei der Partnersprache dennoch relativ hoch ausfallen, genügen sie aber durchaus üblichen Ansprüchen an die Validität in psychosozialen Anwendungsbereichen. Nur bei Verfahren, die einer individuellen Beurteilung dienen sollen, wird nach Lienert eine Validität von $r_{tc} \geqq 0,7$ gefordert[259]. Zu Gruppenvergleichen reichen dagegen sogar schon Werte von $r_{tc} \geqq 0,5$ aus.

Da die Analyse des Beobachtungsverfahrens hinsichtlich der Gütekriterien insgesamt zu zufriedenstellenden Resultaten führt, können die Ergebnisse der gegen Ende des 2. Schuljahres durchgeführten Untersuchung zur Entwicklung der Sprachfähigkeiten am deutsch/italienischen Standort der SESB als ziemlich gesichert angesehen werden.

[259] Lienert 1969, a.a.O., S.310f