25.2.2 Überprüfung der Reliabilität anhand parallelformähnlicher Untersuchungsanordnungen

Außer der Objektivität und inhaltlichen Gültigkeit hat ein Verfahren nachzuweisen, mit welcher Genauigkeit die Daten erhoben werden, ob sich etwa nur zufällige Befunde ergeben oder ob das Verfahren geeignet erscheint, Merkmale hinlänglich stabil und differenzierend zu erfassen. Während sich im Vergleich zu anderen Gütekriterien die Reliabilität von psychometrischen Verfahren ziemlich leicht anhand teststatistischer Rechenmethoden oder auch empirisch, etwa durch die Anwendung eines Retests oder einer Parallelform ermitteln lässt, können die meisten der üblichen Methoden zur Reliabilitätsprüfung des bei der Untersuchung verwendeten Beobachtungsverfahrens leider nicht in Betracht gezogen werden.

Abgesehen davon, dass angesichts der überwiegend auf spontane Sprachäußerungen hin ausgerichteten Untersuchungsanordnungen eine Wiederholung mit denselben Materialien ausschied, war der denkbare Einsatz ähnlicher Bildvorlagen, Spiele oder Texte unter strikter Beibehaltung der Beobachtungskriterien wegen des ohnehin schon recht umfangreichen Programms nicht zu leisten. Doch um wenigstens ansatzweise die Konsistenz des Sprachverhaltens und seine instrumentelle Erfassung durch die Beobachtungskategorien überprüfen zu können, wurde sowohl bei der starken Sprache und der Partnersprache von vornherein für die Bereiche Phonematik/ Prosodie, mündlicher Sprachgebrauch und Hörverständnis eine zweifache Erhebung und Auswertung der Beobachtungen anhand nur leicht modifizierter Kategorien und Bewertungsrichtlinien bei verschiedenen Untersuchungsanordnungen vorgesehen. Da jede Aufgabenstellung aber über den verwandten Beobachtungsbereich hinaus auch der Beobachtung spezifischer Sprachleistungen dienen sollte, handelt es sich nicht um direkt vergleichbare Parallelformen, weshalb auch keine allzu starken Beziehungen, die sich in hohen Reliabilitätskoeffizienten ausdrücken (mindestens 0,70) würden, erwartet werden dürfen.

Da der nach der Paralleltestmethode ermittelte Reliabilitätskoeffizient einer Produkt-Moment-Korrelation zwischen zwei Verteilungen entspricht[252], werden die nach diesem Rechenverfahren ermittelten Interkorrelationen, ergänzt um ihre jeweilige Beziehung zum Gesamtergebnis, angegeben. Bei den an der Untersuchung beteiligten deutschen

[252] Ohne dass Lienert in seinem Standardwerk *Testaufbau und Analyse*, 1969³, explizit auf diesen Zusammenhang hinweisen würde, ergeben sich bei probeweise durchgeführten Vergleichen zwischen der Anwendung der bei ihm auf S.216 für den Reliabilitätskoeffizienten bei Retest- oder Parallelmethode angegebenen Formel 53 und der Berechnung einer Produkt-Moment-Korrelation nach Pearson jeweils identische Ergebnisse.

und italienischen Kindern eines 2. Schuljahres ergeben sich zwischen den parallelformähnlichen Bereichen folgende Korrelationen:

Interkorrelationen und Korrelationen mit Gesamtergebnis (Auszug)

Bezug: max. 25 Punkte, inkl. Schreibfertigkeit	nur starke Sprache / Muttersprache						
	D Kontroll	D SESB	D beide	I Kontroll	I SESB	I beide	D/ I alle
n =	10	18	28	9	18	27	55
Hörverständnis Bild / Textverständnis	0,33	0,10	0,38	0,64	0,47	0,46	0,40
Hörverständnis Bild / Gesamtergebnis	0,70	0,32	0,61	0,49	0,78	0,64	0,62
Textverständnis Bild / Gesamtergebnis	0,78	0,51	0,71	0,55	0,70	0,68	0,66
Pho/Pro Bild / Pho/Pro Lesen	0,30	0,28	0,38	0,65	0,77	0,70	0,53
Pho/Pro Bild / Gesamtergebnis	0,65	0,79	0,74	0,80	0,81	0,68	0,70
Pho/Pro Lesen / Gesamtergebnis	0,78	0,38	0,70	0,59	0,80	0,63	0,66
mündl. Sprachgebrauch: Bild / Textv.	0,95	0,72	0,79	0,96	0,81	0,78	0,80
mündl. Sprach. Bild / Gesamtergebnis	0,82	0,59	0,75	0,69	0,89	0,80	0,77
mündl. Sprach. Textv. / Gesamtergebnis	0,81	0,72	0,71	0,76	0,93	0,90	0,81

Bezug max. 22 Punkte	nur Partnersprache		
	Ps D	Ps I	Ps D/ I
n =	18	18	36
Hörverständnis Bild / Hörverst. Spiel 1	0,50	0,72	0,75
Hörverständnis Bild / Gesamtergebnis	0,65	0,69	0,82
Hörverständ. Spiel 1 / Gesamtergebnis	0,75	0,92	0,87
Pho/ Pro Bild / Pho/ Pro Ps-Lesen	0,22	0,62	0,35
Pho/ Pro Bild / Gesamtergebnis	0,73	0,72	0,87
Pho/ Pro Ps-Lesen / Gesamtergebnis	0,60	0,82	0,59
mündl. Sprachgebrauch: Bild / Spiel 1	0,67	0,50	0,72
mündl. Sprach. Bild / Gesamtergebnis	0,84	0,82	0,91
mündl. Sprach. Spiel 1 / Gesamtergebnis	0,79	0,85	0,89

Anmerkungen:

Die Korrelationen für zusammengefasste Gruppen wurden anhand der Daten der entsprechenden Probanden ermittelt. Da dabei die jeweiligen Beziehungen - gegebenenfalls auch umgekehrte - berücksichtigt werden, ergeben sich anders als bei der Bildung von Mittelwerten zwischen Korrelationen mitunter stärkere oder schwächere Korrelationen, als in Anbetracht der Gruppenkorrelationen zu vermuten wäre.

Bezug: max. 22 Punkte, ohne Schreibfertigkeit	Deutsch	Italiano	D + It.
	D stark + Ps D	I stark + Ps I	alle
n =	46	45	91
Hörv. Bild / Textv.- Hörv. Spiel 1	0,37	0,66	0,68
Hörv. Bild / Gesamtergebnis	0,59	0,64	0,59
Textv.- Hörv. Spiel 1 / Gesamtergebnis	0,72	0,68	0,59
P/P Bild / P/P Lesen - P/P Ps Lesen	0,31	0,70	0,52
Pho/ Pro Bild / Gesamtergebnis	0,73	0,68	0,84
P/ P Lesen - P/P Ps Lesen / Gesamt.	0,60	0,63	0,68
mündl. Sprachg.: Bild/ Textv. - Spiel 1	0,75	0,52	0,60
mündl. Sprach. Bild/ Gesamtergebnis	0,79	0,80	0,86
mündl. Spra.: Textv. - Spiel 1/ Gesamt.	0,73	0,90	0,70

Angesichts der bei allen 3 Werten im Bereich Phonematik/Prosodie erstaunlichen Übereinstimmung bei der italienischen Sprache zwischen den Korrelationen der Gruppe der starken Sprache und der Gruppe Italiano einschließlich der Partnersprache wurden alle Berechnungen mehrfach überprüft.

Die Überprüfung der Reliabilität über die parallelformähnlichen Aufgaben, bei denen die Sprachleistungen nach ähnlichen oder sogar denselben Kriterien bewertet wurden, erweist eine geringere Übereinstimmung als erwartet. Zwar waren von vornherein aufgrund der bei den verwandten Untersuchungsanordnungen gleichzeitig spezifisch zu beobachtenden Sprachbereiche keine allzu starken Zusammenhänge anzunehmen,

aber die meisten Interkorrelationen erfüllen noch nicht einmal die zaghafte Hoffnung auf eine annähernde Beziehung. Lediglich im Bereich „Mündlicher Sprachgebrauch“, bei dem in den Untersuchungsanordnungen des Bildimpuls-gesteuerten Interviews und der Aufgabe zum Textverständnis jeweils Beobachtungen zum Wortschatz und zur Morphosyntax ausgewertet wurden, ergeben sich bei der starken Sprache sowohl im Deutschen als auch im Italienischen relevante Korrelationen zwischen r = 0.72 und r = 0.96. Ansonsten fallen die Interkorrelationen je nach Sprache oder Lerngruppe recht unterschiedlich aus, so dass die anhand der parallelformähnlichen Aufgaben festgestellten Zusammenhänge kaum auf eine allgemeine Zuverlässigkeit dieser Untersuchungsanordnungen hinweisen.

Dennoch sind die Ergebnisse sehr aufschlussreich. Einerseits lassen sie erkennen, dass zwischen den Anforderungen beim Textverständnis und dem Hörverständnis im freien Gespräch sowie hinsichtlich der phonetisch-prosodischen Sprachebene beim Lesen und im Gespräch offensichtlich derart große Unterschiede bestehen, dass die vermeintliche Ähnlichkeit der Beobachtungskategorien kaum zum Tragen kommt. Vor allem aber weisen die auffälligen Differenzen auf sprachstruktur- und lernvoraussetzungsabhängige Unterschiede hin, womit der Anspruch auf Reliabilität bei einem Verfahren zur Beobachtung von bilingualen Lernprozessen relativiert wird. Wie auch hinsichtlich der Schwierigkeit und der Trennschärfe verlangt ein Verfahren zur Erfassung bilingualer Lernprozesse statt der Ermittlung einer Gesamtreliabilität, bei der bestehende Unterschiede sowieso oft rechnerisch ausgeglichen werden, die Überprüfung der Reliabilität getrennt nach Sprache und Beobachtungsgruppe. Es gilt demnach bei jeder Sprache und jeder Lerngruppe gesondert zu überprüfen, ob das Verfahren für diese Zielgruppe als zuverlässig betrachtet werden kann.

Bevor die zielgruppenorientierte Reliabilitätsüberprüfung in Angriff genommen wird, soll deren Erfordernis noch anhand eines bei den Interkorrelationen besonders auffälligen Unterschiedes zwischen den Sprachen verdeutlicht werden. Zwischen den Beobachtungen im Bereich Phonematik/Prosodie beim Interview und beim Lesen scheinen sowohl bei der Lerngruppe der starken Sprache als auch jeweils bei der Partnersprachengruppe im Deutschen kaum relevante Zusammenhänge (r = 0.22 bis r = 0.38) zu bestehen, im Italienischen aber durchaus (r = 0.62 bis r = 0.77). Daher ist anzunehmen, dass die Fähigkeit einer angemessenen Intonation und normgerechten Aussprache beim Lesen unabhängig von der Lesefertigkeit, die Schülern eines 2. Schuljahres anscheinend teilweise noch arge Probleme bereitet, bei der italienischen Sprache von ihrer gegenüber der deutschen Sprache stärkeren phonem-graphemischen Übereinstimmung unterstützt wird. Die Differenz lässt sich in diesem Fall eher mit dem sprachstrukturellen Unterschied erklären als mit ebenfalls denkbaren

unterschiedlichen Schwierigkeitsgraden der in den beiden Sprachen vorgelegten Texte, denn die Interkorrelationen unterscheiden sich sowohl bei den literarischen Textvorlagen in der starken Sprache als auch bei den Einzelsätzen, die zur Beobachtung der Lesefertigkeit in der Partnersprache benutzt wurden, in etwa gleichem Maße.

25.2.3 Zielgruppenorientierte Reliabilitätsprüfung

Zur Einschätzung der Zuverlässigkeit der mit der Pilotfassung des Beobilingua-2dit-Verfahrens erhobenen Beobachtungsdaten werden Reliabilitätsberechnungen sowohl für jede einzelne Sprach- und Leistungsgruppe als auch unter Zusammenfassung der Gruppen nach Sprachstatus oder Sprache ausgeführt. Gruppenweise verschiedene Berechnungen scheinen nicht nur angeraten, weil sich die Korrelationsmatritzen in manchen Bereichen auffällig nach Sprache und/oder Leistungsgrad unterscheiden, sondern auch wegen des Anwendungsbereichs bei einer früh einsetzenden bilingualen Erziehung, weshalb gewisse Unterschiede von der Konstruktion her beabsichtigt sind.

Das Verfahren soll zwar die Varianz der Sprachkompetenzen in den verschiedenen Spracherwerbsstadien vom Anfänger bis zur altersgemäßen Sprachbeherrschung über die ganze Bandbreite zuverlässig erfassen, aber insbesondere der differenzierten Beobachtung von Leistungsunterschieden im Bereich der sich erst noch entwickelnden Sprachfähigkeit in einer schwächer ausgebildeten simultan erworbenen Erstsprache oder einer Zweitsprache dienen. Dabei soll das Instrumentarium eine altersgemäße, monolingual vergleichbare Sprachkompetenz nicht völlig außer Acht lassen, aber in stärkerem Maße verschiedene Stufen des sich diesem Ziel nähernden Spracherwerbs unterscheiden. Daher werden an die Reliabilität zwecks Erfassung einer größeren Leistungsvariabilität in dem Bereich der schwächeren Sprachkompetenzen höhere Anforderungen gestellt als bei den annähernd altersgemäßen oder muttersprachlichen Sprachkompetenzen.

Bei der statistischen Prüfung der Reliabilität können aufgrund der einer heterogenen Testbatterie ähnelnden Konstruktion einige der dafür entwickelten Methoden auch auf das Beobilingua-2dit-Verfahren übertragen werden. Wenn ein Verfahren nur einmal dargeboten wurde und somit weder Daten aus einer Wiederholung, Parallelform oder Verfahrenshalbierung vorliegen, kann die Reliabilität durch die Analyse der inneren Konsistenz geprüft werden. Der ermittelte Reliabilitätskoeffizient ist dann ein Konsistenzkoeffizient und kennzeichnet die instrumentale Reliabilität unabhängig von den Durchführungsbedingungen.

Für eine Konsistenzanalyse haben vor allem Kuder und Richardson diverse Formeln entwickelt, die jedoch nur unter der Bedingung angewendet werden dürfen, dass alle Aufgaben eines Verfahrens dieselben Faktoren erfassen.[253] Daher kann bei dem an der SESB eingesetzten heterogenen Beobachtungsverfahren die Gesamtreliabilität des Verfahrens nicht mit einer Kuder-Richardson-Formel ermittelt werden, wohl aber die Konsistenz einzelner Untersuchungsbereiche, bei denen sich die Beobachtungs- und Auswertungskriterien jeweils auf denselben sprachlichen Aspekt (die Phonetik, das Hörverständnis, die Syntax, den Wortschatz, usw.) beziehen. Die Anwendung der Formula 21 führt allerdings nur zu einem groben, meist weit unter dem eigentlichen Reliabiliätswert liegenden Schätzwert, weshalb die Koeffizienten auch nur in Bezug auf die italienische Sprache bestimmt wurden, um anhand der verfahrensinternen Näherungswerte Entscheidungshilfen hinsichtlich der später vorzunehmenden Straffung und Überarbeitung des Verfahrens zu gewinnen.

Konsistenzkoeffizienten Italienische Sprachkompetenz (starke Sp. und Ps)			
Hörverständnis - Bild	0,56	Lesefertigkeit	- 0,73
Phonematik/Prosodie – Bild	0,40	Phon./Prosodie – Lesen	0,00
Wortschatz – Bild	0,38	Konzepte	0,62
Morphosyntax – Bild	0,81	Schreibfertigkeit (nur st. Spr.)	- 0,68
mündl. Sprachgebrauch - Bild	0,82	mündl. Sprach. Textv. / Spiel 1	0,59
Hörverständnis Spiel 1 (nur Ps)	0,24	mündl. Sprachgebrauch gesamt	0,62
Hörverständnis gesamt	0,20	kommunikatives Sprachverh.	0,33

Die Untersuchungsbereiche Lese- und Schreibfertigkeit, die sich bereits als nicht trennscharf erwiesen, fallen auch durch besondere Inkonsistenz auf.

Für die Einschätzung der Gesamtreliabilität von heterogenen Testbatterien hat Mosier 1943 eine Formel entwickelt, die auch die Interkorrelationen von Untertests – wie sie im Beobilingua-2dit-Verfahren den einzelnen Untersuchungsbereichen entsprechen – berücksichtigt, und daher keine Homogenität voraussetzt. Da diese Formel außerdem noch die Gewichtsfaktoren der Untertests einbezieht, erfordert ihre Anwendung allerdings umfangreiche Rechenoperationen.

[253] Quelle siehe Fußnote 215. – Die Anwendung der KUDER-RICHARDSON-Formula 21, die nur mit den statistischen Kennwerten (Mittelwert und Varianz) auskommt, setzt außerdem voraus, dass auch gleiche Schwierigkeiten vorliegen. Wenn eine oder gar beide Bedingungen verletzt werden, ist die mit dieser Formel bestimmte Konsistenz nur äußerst ungenau.

Gesamtreliabilität des Verfahrens nach Zielgruppen (nach Formel 128) [254]

Partnersprache Deutsch	0,83
Partnersprache Italienisch	0,90
beide Partnersprachen zusammen	0,89
starke Sprache Italienisch	0,94
starke Sprache Deutsch	0,87
starke Sprachen zusammen	0,92
insgesamt (It + D, stark + Ps)	0,91

Bei diesen hohen Reliabilitätskoeffizienten darf davon ausgegangen werden, dass die Sprachkompetenzen der SESB-Schüler gegen Ende des 2. Schuljahres zuverlässig beobachtet wurden. Die jeweils etwas niedrigere Reliabilität in Bezug auf die deutsche Sprache erklärt sich durch die allgemein stärker entwickelte Kompetenz im Deutschen. Bei der Konstruktion des Verfahrens wurde eine Differenzierung eher im schwächeren als im oberen Leistungsbereich beabsichtigt. Insofern ist es jedoch erstaunlich, dass das Verfahren anscheinend bei den Lerngruppen der starken Sprache jeweils reliabler zu sein scheint als bei den Lerngruppen der Partnersprache, wobei dieses Ergebnis in Bezug auf die italienische Sprache vermutlich darauf zurückzuführen ist, dass an der SESB auch Schüler der italienischen Sprachgruppe zugeordnet wurden, deren

[254] Formeln 128 und 129 bei Lienert 1969, a.a.O., S. 380 und 381. – Die zur Ermittlung der Reliabilität des Verfahrens benutzten Formeln von Mosier sind weniger verbreitet als die Kuder-Richardson-Formeln und werden daher in der vereinfachten Schreibweise von Lienert auf Tabellenblatt 24 wiedergegeben. – Unter Zusammenfassung aller in beiden Partnersprachen Italienisch und Deutsch dargebotenen Untersuchungsbereiche erfolgte die Ermittlung der Gesamtreliabilität des Verfahrens für die Lerngruppe Partnersprache zunächst nach der Gewichtsformel von Moisier mit einem Ergebnis von $bat^{r}tt = 0{,}88$ (Berechnung siehe Tabellenblatt 14). Anschließend wurden dieselben Untersuchungsbereiche ohne Berücksichtigung der Gewichte mit der durch Lienert vereinfachten Formel 128 überprüft, was zu einem fast ähnlichen Ergebnis von $bat^{r}tt = 0{,}89$ führte (siehe Tabellenblatt 15). Aufgrund der nur geringfügigen Differenz wurde bei den weiteren Reliabilitätsberechnungen vom Gebrauch der Formel 129 abgesehen und die kleine Einbuße an Präzision durch die Vernachlässigung der Gewichte wegen der wesentlich kürzeren Rechenprozedur in Kauf genommen. – Auf Tabellenblatt 15 werden jeweils nur die Ergebnisse und die zuvor separat errechneten Werte angegeben, die in die Formel einzusetzen sind, damit die Berechnung nachvollziehbar ist. Auf die Wiedergabe der vollständigen Berechnungen wird indessen verzichtet, um den vorliegenden Bericht nicht durch das statistische Datenmaterial anschwellen zu lassen. Außerdem wird in ähnlicher Weise zum Vergleich für jede Lerngruppe auch der aus einer Berechnung nach der Formula 21 resultierende Reliabilitätskoeffizient angegeben, der jeweils niedriger ausfällt, was keineswegs der anhand der Formel 128 errechneten Reliabilität des Verfahrens widerspricht, da das Formula 21-Verfahren bei einem heterogenen Verfahren nicht die angemessene statistische Methode darstellt und außerdem gerade in Bezug auf die Partnersprache, also den bei dieser Untersuchung besonders interessierenden Lernbereich, als einen unteren Schätzwert durchaus befriedigende Resultate zeigt.

sogenannte starke Sprache noch nicht altersgemäß entwickelt ist. Mit sehr großer Zuverlässigkeit hat das Verfahren genau diese bei der nichtdeutschen Sprachgruppe auftretende Sprachvarianz bestätigt.

In Bezug auf die deutsche Sprache ergibt sich der höhere Reliabilitätskoeffizient im starken Bereich jedoch vorwiegend algebraisch, weil in die Formel 128 die Summe der Trennschärfekoeffizienten einzusetzen ist (Tabellenblatt 15, jeweils Spalte 1). Da diese bei der deutschen Sprache aber aufgrund der groben Berechnungsmethode bei verschiedenen Untersuchungsanordnungen recht niedrig ausfielen, bei der starken Sprache aber mehr Bereiche mit zum Teil höheren Trennschärfen einzubeziehen sind, fällt der Koeffizient als Resultat der Rechenoperation zwangsläufig höher aus. Auch wenn sich die Differenz dadurch hinreichend erklären lässt und die Gesamtreliabilität auch für die deutsche Partnersprache durchaus zufriedenstellend ist, sollte dennoch durch eine Überarbeitung der für die Partnersprache relevanten Kategorien (z.B. Hörverständnis und Morphosyntax) versucht werden, die Zuverlässigkeit für diese Zielgruppe noch zu erhöhen.

Auch in Bezug auf die starke Sprache könnte die Gesamtreliabilität sicher noch durch einen anderen Bewertungsmaßstab dieser Bereiche erhöht werden. Außerdem bietet sich an, wenig konsistente Bereiche wie die Lese- und Schreibfertigkeit zu eliminieren. Da diese Bereiche aber wiederum eigenständige Faktoren der Sprachkompetenz erfassen, ohne deren Einbezug eine Sprachstandserhebung unvollständig und weniger valide ist, sollte nicht unbedacht auf wesentliche Bereiche zur Erhöhung der Reliabilität verzichtet werden. Nach Lienert können heterogene Verfahren trotz geringer Konsistenz eine höhere praktische Validität besitzen als sehr homogene Verfahren, die aufgrund der homogenen Aufgaben eine hohe Reliabilität aufweisen. Zwischen diesen beiden Güteeigenschaften besteht eine partielle Inkompatibilität, „*indem man das eine anstrebt, gefährdet man das andere.*“[255] Angesichts der nicht aufzulösenden Antinomie von Validität und Reliabilität ist abzuwägen, welcher Güteeigenschaft des Verfahrens der Vorzug zu geben ist. Meines Erachtens könnten Beobachtungen zur motorischen Schreibfertigkeit, die sich in beiden Sprachen als besonders inkonsistent erwiesen hat, ohne gravierende Einbußen an der Validität entfallen, aber zur Ergänzung mündlicher Aspekte der Sprachfähigkeit sollten wenigstens in einem der Bereiche Schreiben oder Lesen Beobachtungen trotz der im 2. Schuljahr noch mangelnden Konsistenz beibehalten werden. Bei der Konstruktion eines Verfahrens, das die Diagnose eines komplexen Verhaltens wie Sprache beabsichtigt, sind zwecks Ausgewogenheit Kompromisse unvermeidlich.

[255] Lienert 1969, a.a.O., S.294