Webergänzung zu Kapitel 6

6.4.3 Die Beurteilung des Tests

Das Kapitel „Tests“ wurde im Buch in den wichtigsten Zügen dargestellt, wobei der Blick vor allem auf die Anwendung von Sprachtests in einer experimentellen Untersuchung gerichtet war. Hier ging es insbesondere um die Frage, ob der Test geeignet ist, die Fortschritte von Lernenden zwischen einem „Vorher-“ und einem „Nachher“-Zustand zu ermitteln. Aber ganz unabhängig davon, wofür man den Test einsetzen will, kann es wichtig sein, die Qualität von einem Test beurteilen zu können. Das kann man sowohl bei geschlossenen als auch bei offenen Tests machen.

6.4.3.1 Einheitenanalyse (*item analysis*) bei geschlossenen Tests

Wenn man seinen Test so gut wie irgend möglich entworfen haben möchte, muss man sich jeden einzelnen Teil bzw. jede einzelne Einheit des Tests separat ansehen. Man muss also, anders ausgedrückt, eine Einheitenanalyse durchführen, mit der man entscheidet, ob eine bestimmte Einheit gut genug ist in dem Sinne, dass eine korrekte Antwort wahrscheinlich einen Überblick darüber geben wird, was der Getestete weiß oder was er kann. Diese Art von Einheitenanalyse wird normalerweise gemacht, **nachdem** ein Test durchgeführt wurde und die Ergebnisse vorhanden sind. Natürlich muss auch eine Art von Einheitenanalyse durchgeführt werden, **bevor** ein Test durchgeführt wird. Das bedeutet ganz einfach, dass man sorgfältig alle Einheiten überprüfen sollte, bevor man den Test verwendet, und sich dabei fragen sollte, ob jede Einheit wirklich repräsentativ ist für das, was man testen möchte, und ob sie nicht zu leicht oder zu schwer ist. Eine Einheit ist wahrscheinlich zu schwierig, wenn man das Gefühl hat, dass selbst gute Lerner, also Lerner mit viel Wissen oder solche, die viele Stunden geübt haben, sie nicht richtig beantworten könnten. Gleichermaßen ist eine Frage, die allein schon auf der Grundlage von gesundem Menschenverstand beantwortet werden kann, wahrscheinlich keine gute Frage, es sei denn, man testet den „gesunden Menschenverstand“. Es ist jedoch nicht möglich, sich sicher zu sein, ob ein Test wirklich ein guter Test ist und ob die Einheiten gut gewählt sind, bevor man nicht die Daten von Personen hat, die ihn tatsächlich absolviert haben.

Betrachten wir zunächst Multiple-choice-Tests[1] als Beispiel für die Überlegungen, die bei Einheitenanalysen durchgeführt werden müssen. Ein Multiple-choice-Test besteht, wie andere Tests auch, aus mehreren Einheiten

---

[1] In manchen Veröffentlichungen nennt man Multiple-choice-Tests, bei denen jeweils nur eine Antwortmöglichkeit richtig ist, Single-choice-Tests. Wir bleiben bei der eingeführten Terminologie.

(Fragen). Einige dieser Einheiten können „besser“ als andere sein. Zum Beispiel kann eine bestimmte Einheit sehr schwierig sein in dem Sinne, dass praktisch niemand die richtige Antwort ankreuzt, oder so einfach, dass jeder die richtige Antwort ankreuzt (vielleicht weil die Distraktoren so unwahrscheinlich sind, dass niemand sie auswählt). Es kann auch passieren, dass eine Einheit etwas anderes testet als die anderen Einheiten oder dass der Test als Gesamtheit nicht das testet, was man testen möchte. Uns geht es in diesem Zusatzmaterial vor allem um die logischen und rechnerischen Verfahren, die man zur Überprüfung der Qualität von Tests anwenden kann.[2]

Es hat einen großen Einfluss auf die Schwierigkeit des Tests, ob immer nur eine der vorgegebenen Antworten richtig sein kann, oder ob keine, eine, zwei, drei bis alle Antworten richtig sein können. Im ersten Fall würde bei drei Distraktoren die Ratewahrscheinlichkeit bei 25 % liegen (1:4), im zweiten Fall wäre bei vier vorgegebenen Antworten die Ratewahrscheinlichkeit 1:16, 6,25 %. Und wenn man mehr als vier Antworten vorgibt, wird die Ratewahrscheinlichkeit noch geringer.

Eine Einheit kann sehr leicht sein. Wenn in einem Multiple-choice-Test alle Getesteten die korrekte Antwort auswählen, haben wir ein 100%-korrekt-Ergebnis. Das niedrigste Prozent-korrekt-Ergebnis wird durch die Wahrscheinlichkeit, dass jeder eine der Optionen zufällig ankreuzt, zum Beispiel durch Raten, bestimmt.[3] Wenn es vier Möglichkeiten gibt und den Getesteten bekannt ist, dass nur eine Antwort anzukreuzen ist (es gibt eine korrekte Antwort und drei Distraktoren), beantworten z.B. auch Lerner, die kein Arabisch sprechen, bei einem solchen Test zur arabischen Sprache wahrscheinlich 25 % der Testfragen korrekt; wenn es drei Möglichkeiten gibt, haben wir um die 33 % korrekte Antworten. Falls überhaupt keine Möglichkeiten gegeben werden (also im Falle von offenen Fragen), sollte im Falle des Ratens das Prozent-korrekt-Ergebnis bei 0 % liegen.

Wann ist eine Einheit schwierig oder leicht? Je mehr Getestete eine Einheit korrekt beantworten, desto leichter ist sie. Das gilt sowohl für geschlossene als auch für offene Tests. Der Grund muss kein schlechter Test sein; vielleicht gibt es so viele richtige Antworten, weil die Getesteten sich sehr gut auf den Test vorbereitet haben oder weil sie besonders intelligent sind oder weil sie eine exzellente Lehrperson hatten. Andere Gründe wären, dass die Distraktoren in einem Multiple-choice-Test nicht attraktiv genug waren oder dass die Einheit korrekt zu beantworten nichts weiter als gesunden Menschenverstand, also nicht die zu testenden Sprachkenntnisse, erforderte. In den letzten beiden Fällen ist die Einheit nicht gut konstruiert.

---

2 Siehe hierzu: R. E. Ebel, 1972. *Essentials of Educational Measurement*. Engelwood Cliffs, New York: Prentice-Hall.

3 Sehr häufig sind nicht alle Distraktoren gleich attraktiv oder plausibel. Wenn zwei der vier Optionen sofort von der Hand gewiesen werden können, wird die Wahrscheinlichkeit, die richtige Antwort anzukreuzen, drastisch höher.

proportion-Wert (*p*-Wert)

Der Anteil der Personen, die eine Einheit korrekt beantworten, wird *p*-Wert (von *proportion*) genannt.[4] Wenn 90% aller Getesteten die korrekte Antwort geben, ist der *p*-Wert dieser Einheit 0,90. Der maximale *p*-Wert ist 1 (was der Fall ist, wenn jeder die korrekte Antwort weiß), und der minimale ist 0 (wenn niemand die richtige Antwort weiß). Meist versucht man, einen *p*-Wert zwischen 0,50 und 0,75 zu erreichen, obwohl es natürlich Gründe geben kann, eine Einheit in einem Test zu haben, deren *p*-Wert in diesem Fall größer als 0,75 wäre. Zum Beispiel kann man die erste Frage in einem Test mit Absicht leicht gestalten, um bei den Prüfungsteilnehmern Stress abzubauen. Es ist nicht wirklich sinnvoll, besonders schwierige Fragen in geschlossene Tests aufzunehmen, in der Absicht, die hervorragenden von den besonders guten Lernenden unterscheiden zu können, denn die hervorragenden Lernenden werden (fast) alle Fragen richtig beantworten können, die guten werden trotzdem einige Fehler machen, und sogar die schlechten Lernenden werden durch Raten einige Fragen richtig beantworten.

distractor-Wert (*d*-Wert)

Wenn zu viele Personen die falsche Antwort geben, z.B. indem sie alle denselben Distraktor statt der korrekten Antwort auswählen, kann auch etwas an der Einheit falsch sein. Man kann die Anzahl der Getesteten berechnen, die einen bestimmten Distraktor gewählt haben. Dieser Wert, den wir den *d*-Wert nennen (*d* steht hier für Distraktor), sollte kleiner als der *p*-Wert der dazugehörigen Einheit sein. Wenn er höher ist, dann ist wahrscheinlich etwas mit der Einheit nicht in Ordnung. Man könnte natürlich denken, es sei besonders schlau, einen naheliegenden Irrtum als Distraktor zu wählen, aber in der Testtheorie sieht man das nicht so, alle Distraktoren sollten mit einer ähnlichen Wahrscheinlichkeit gewählt werden und die richtige Antwort sollte häufiger gewählt werden als die Distraktoren.

Neben dem *p*-Wert einer Einheit ist wichtig, wie gut diese Einheit zwischen guten und schlechten Lernenden differenziert. Diese Differenzierung wird der *D-Index* genannt. Idealerweise würde man erwarten, dass das Durchschnittsergebnis derjenigen, die eine bestimmte Einheit korrekt beantworten, besser ist als das Durchschnittsergebnis derjenigen, die sie falsch beantworten. Wenn dies nicht der Fall sein sollte, befindet man sich in der unbequemen Lage, beantworten zu müssen, warum eine bestimmte Einheit von schlechteren Lernenden richtig beantwortet wird, während gute Lernende sie falsch beantworten.

---

[4] Damit der *p*-Wert eine sinnvolle Information liefert, sollte die Anzahl der Getesteten größer als 25 sein. Dieser *p*-Wert sollte nicht mit dem *p* aus dem Bereich der Signifikanz (von *probability*) verwechselt werden.

Tabelle 1: Beispiel eines Sprachtests

beste Schüler		schlechteste Schüler	
Schüler	Ergebnis bei Einheit x	Schüler	Ergebnis bei Einheit x
1	1	16	1
2	1	17	0
3	0	18	1
4	1	19	0
5	1	20	0
Summe	4	Summe	2
p-Wert	0,80	*p*-Wert	0,40

D-Index

Die Berechnung des D-Indexes ist recht einfach. Nehmen wir an, dass wir eine Gruppe von 20 Testteilnehmern haben und herausfinden möchten, ob Einheit x gut genug zwischen guten und schlechten Lernenden differenziert. Dazu gehen wir wie folgt vor:

Wir bringen die Ergebnisse aller Getesteten in eine Rangordnung von gut nach schlecht. Man nimmt die besten 25 % und die schlechtesten 25 %, d.h. in unserem Fall die fünf besten und die fünf schlechtesten Schüler. Tabelle 1 sagt uns, ob die fünf besten Schüler (1 bis 5) und die fünf schlechtesten Schüler (16 bis 20) eine bestimmte Einheit richtig oder falsch beantwortet haben (1 = richtig, 0 = falsch). Man berechnet in jeder Gruppe den Anteil der Schüler, die die Einheit korrekt beantwortet haben (hier liegen die *p*-Werte bei 0,80 und 0,40), und subtrahiert den *p*-Wert der Gruppe der schlechten von dem der guten Schüler. Der erzielte Wert ist der D-Index, in unserem Fall 0,80 – 0,40 = 0,40.

Wir können sehen, dass der D-Index zwischen +1 und –1 schwanken kann, wobei ersteres eine perfekte Differenzierung und letzteres eine Einheit mit einer völlig falschen Differenzierung zwischen guten und schlechten Schülern bedeutet. Ein D-Index von ungefähr 0 bedeutet, dass die Einheit überhaupt nicht zwischen guten und schlechten Schülern differenziert. In diesem Fall sollte man sich überlegen, die Einheit aus dem Test zu entfernen.

Wenn ein Test aus vielen verschiedenen Teilfragen besteht, ist es unpraktisch, für jede Teilfrage „von Hand“ auszurechnen, ob sie zu den anderen passt. Es ist sinnvoller, dafür ein Statistikprogramm zu benutzen. Um die interne Konsistenz eines Tests zu prüfen, kann man zum Beispiel mit einer Reliabilitätsanalyse in SPSS sehr einfach das Cronbach Alpha ausrechnen lassen. Das Cronbach Alpha gibt dann an, wie hoch die Verlässlichkeit eines Tests insgesamt ist, wobei für jede Teilfrage berechnet werden kann, wie gut sie mit den anderen korreliert. Das ist der D-Index.

Verlässlichkeit

6.4.3.2 Offene Tests

In offenen Tests muss der Lernende die Antwort selbst formulieren. Mündliche Tests sind dafür sehr typisch und eignen sich nicht besonders gut für statistische Analysen, weil jeder mündliche Test ein einzigartiges Gespräch mit einer individuellen Interaktion zwischen Prüfer und Prüfling darstellt. Diese Einzigartigkeit macht es unmöglich, einen bestimmten Test mit den Tests anderer Schüler zu vergleichen, weil man keine Einheitenanalyse anwenden oder die Verlässlichkeit berechnen kann, wie dies in geschlossenen Tests möglich ist. Es wurden Versuche unternommen, selbst Prüfungsgespräche so zu strukturieren, dass eine objektive Bewertung möglich ist – denn objektive Bewertung ist die erste Notwendigkeit, um Tests zu analysieren –, aber die Ergebnisse dieser Versuche sind in der Praxis nicht immer überzeugend.

Einige offene Tests jedoch sind in gewisser Weise geschlossenen Tests sehr ähnlich. Ein Beispiel ist ein Grammatiktest, in dem der Schüler Sätze vervollständigen muss. Nehmen wir an, der Schüler soll einen Satz vervollständigen, der mit

Kaum jemals _____

beginnt, und der Forscher möchte wissen, ob der Schüler weiß, dass nach dieser Wendung eine Inversion von Subjekt und Prädikat stattfindet. Das Ergebnis kann mit 1 (korrekt) bewertet werden, wenn der Schüler folgendes schreibt:

Kaum jemals kam er im Abend ans Haus

weil „kam er“ die geprüfte Inversion darstellt. Dass der Schüler ansonsten einige Fehler gemacht hat, wäre hier irrelevant. Das Ergebnis wäre jedoch 0 (falsch), wenn der Schüler:

Kaum jemals er kam vor Mitternacht nach Hause

schreibt, weil die Inversion nicht erfolgt ist. Bei einem offenen Test mit dieser Art von Einheiten kann die Umsetzung in Zahlen ebenso wie bei geschlossenen Fragen erfolgen.

Wenn eindeutige Korrekturvorschriften verwendet werden, sind diese Tests wie geschlossene auszuwerten, und man kann die Verlässlichkeit ebenfalls mit einem Reliabilitätstest zur Berechnung des Cronbach Alpha überprüfen.

6.4.3.3 Gemischte Tests

Bei Sprachtests ist es nicht ungewöhnlich, Testreihen zu benutzen, die aus geschlossenen und offenen Tests bestehen. Solch eine Reihe kann zum Beispiel aus zwei geschlossenen Tests (einem Vokabeltest mit 60 Einheiten und

einem Satzergänzungstest mit 100 Einheiten) und zwei offenen Tests (einem Aufsatz und einer mündlichen Prüfung) bestehen. Die Endnote des Schülers, der alle vier Tests absolviert hat, kann der Mittelwert aus allen vier Tests, dargestellt in Prozenten, sein, oder sie kann das Ergebnis einer anderen Art von Berechnung sein. Man kann beispielsweise einen Testteil stärker gewichten wollen, beispielsweise wenn man die Note für den mündlichen Test als wichtiger erachtet als die Note für den Vokabeltest. Für die beiden geschlossenen Tests kann man die Verlässlichkeit schätzen; dies wäre bei den offenen Tests sinnlos. Man kann allerdings die Korrelationen zwischen den vier Einzeltests berechnen. Diese Korrelationen wären aber wahrscheinlich nicht sehr hoch.

6.4.3.4 Validität und Verlässlichkeit

Im Buch wurde bereits auf diese Gütekriterien bei Tests eingegangen, hier soll noch ein Rechenverfahren zur Messung der Verlässlichkeit ergänzt werden.

Verlässlichkeitskoeffizient

Wenn man parallele Tests benutzt, kann man den Korrelationskoeffizienten (der hier auch der Verlässlichkeitskoeffizient genannt wird; mehr zu Korrelationen lesen Sie in Kapitel 9) von zwei parallelen Tests berechnen. Beispielsweise kann man zwei Grammatiktests konstruieren, die sich sehr ähneln und dieselbe Anzahl von Einheiten (sagen wir: 100) mit derselben Aufteilung in Kategorien (zum Beispiel 10 Einheiten Wortreihenfolge, 8 Einheiten Präsensformen unregelmäßiger Verben etc.) enthalten und in denen jeder Einheit des einen Tests eine Einheit mit demselben Schwierigkeitsgrad im anderen Test entspricht. Hier sollte der Verlässlichkeitskoeffizient bei 0,85 oder höher liegen. Obwohl man meinen könnte, dass es einfach ist, einen parallelen Test zu entwerfen, ist es in der Praxis dann doch sehr schwierig, zwei Tests vollkommen parallel zu gestalten.

Weil sowohl Doppel- als auch parallele Tests ihre Nachteile haben, werden normalerweise Tests der internen Stimmigkeit benutzt, um die Verlässlichkeit einzuschätzen.

Das Split-half-Verfahren misst die interne Stimmigkeit und wird durchgeführt, indem man einen Test in zwei Hälften aufteilt. Diese Hälften können aus der ersten und zweiten Hälfte des Tests bestehen oder, und dies ist meist besser, jeweils aus den ungeraden und den geraden Einheiten des Tests. Auf diese Weise kann man die Korrelation zwischen den Ergebnissen der beiden Testhälften errechnen. Der errechnete Korrelationskoeffizient gibt einem die Verlässlichkeit für den halben Test. Um den Verlässlichkeitskoeffizienten für den gesamten Test (r_k) zu erhalten, muss man die Spearman-Brown-Formel benutzen:

$$r_k = \frac{2r_1}{r_1 + 1}$$

in der r_1 der Korrelationskoeffizient ist, den man erhält, wenn man die zwei Testhälften korreliert.

Ein Beispiel: Nehmen wir an, wir haben einen Grammatiktest, der aus 100 Einheiten besteht, durchgeführt. Wir teilen nun den Test in zwei Hälften, jede mit 50 Einheiten. Nun berechnen wir die Korrelation zwischen den zwei Testhälften. Nehmen wir an, die Korrelation (r_1) ist 0,86. Setzen wir dies in die Formel ein, so erhalten wir:

$$r_k = \frac{(2)\,(0{,}86)}{0{,}86+1} = \frac{1{,}72}{1{,}86} = 0{,}92$$

Ein Verlässlichkeitskoeffizient von 0,92 zeigt uns, dass der Test sehr verlässlich ist.

Man kann die Verlässlichkeit eines Tests außerdem noch mit Hilfe des Cronbach Alpha oder der KR (Kuder und Richardson)-21-Formel einschätzen. Da beide Verfahren etwas komplexer sind, gehen wir hier nicht darauf ein, sondern verweisen nur darauf, dass es inzwischen recht viele Möglichkeiten gibt, die Verlässlichkeit eines Tests zu prüfen. Die meisten Statistikprogramme bieten gute Verlässlichkeitstests, so dass man dies selber nicht berechnen muss.

Die Verlässlichkeit eines Tests wird von einigen Faktoren beeinflusst, manche davon sind bereits im Buch angeführt, hier geben wir eine umfangreichere Liste:

- Testlänge (je länger ein Test ist, desto verlässlicher ist er)
- Zusammensetzung der Gruppe der Getesteten (wenn alle Schüler praktisch die gleiche Wissensbasis haben, gleich klug sind und ungefähr das gleiche Ergebnis erzielen, ist die Verlässlichkeit niedrig)[5]
- Zeit, die für den Test zur Verfügung steht (wenn die Schüler nicht genug Zeit zur Verfügung haben, ist die Verlässlichkeit im Allgemeinen niedrig)
- Homogenität der Einheiten (wenn die Einheiten den gleichen Aspekt testen, ist die Verlässlichkeit höher, als wenn sie dies nicht tun)
- Objektivität der Bewertung (diese ist bei Multiple-choice-Tests normalerweise gegeben, aber selten bei offenen Tests)
- D-Index der Einheiten (wenn die Einheiten gut zwischen gut und schlecht differenzieren, ist die Verlässlichkeit höher, als wenn sie dies nicht tun).

Dabei ist der wichtigste Faktor die Testlänge. Es ist tatsächlich so, dass man, wenn man eine nicht besonders hohe Verlässlichkeit bei einem Test hat (etwa 0,50) und man gerne eine Verlässlichkeit von 0,80 erreichen möchte, den Test

[5] Dass das so ist, kann man leichter nachvollziehen, wenn man an die Konsequenzen einer geringen Varianz für eine eventuelle Testwiederholung denkt. Wenn alle Ergebnisse ganz dicht beieinander liegen, kann es gut sein, dass bei einer Wiederholung diejenigen schlechter abschneiden, die vorher besser abgeschnitten haben.

viermal so lang machen muss. Eine Berechnung der benötigten Testlänge zum Erzielen eines bestimmten Verlässlichkeitskoeffizienten – von zum Beispiel 0,90 statt der bisher erreichten 0,50 – erfordert die Benutzung der Spearman-Brown-Korrekturformel. Nehmen wir also an, ein Test hat 30 Einheiten und einen Verlässlichkeitskoeffizienten von 0,50.

Die Formel, die hierbei anzuwenden ist, lautet:

$$M = \frac{r_a}{r_o}\left(\frac{r_o - 1}{r_a - 1}\right)$$

wobei

M für den Multiplikationsfaktor steht und

r_a für die Verlässlichkeit, die man mit dem Test erreichen möchte (in diesem Fall 0,90), und

r_o die Verlässlichkeit ist, die der Test in seiner jetzigen Länge hat (bei uns 0,50).

Setzen wir diese Werte in die Formel ein, so erhalten wir:

$$M = \frac{0,90}{0,50}\left(\frac{0,50-1}{0,90-1}\right) = (1,8)\,(5) = 9$$

Dies bedeutet, dass der Test 9-mal so lang sein muss, um einen Verlässlichkeitskoeffizienten von 0,90 zu erreichen. Unser Test hatte 30 Einheiten; wir müssten ihn also auf 270 Einheiten aufblähen (und dabei beachten, dass die zusätzlichen Einheiten die gleichen Aspekte testen …). Es ist offensichtlich, dass ein derart langer Test andere Probleme wie beispielsweise Erschöpfung bei den Schülern verursacht, die dann wiederum die Verlässlichkeit senken würden.

Aufgaben

1. Nehmen Sie die folgenden Ergebnisse eines Grammatiktests und das Ergebnis jedes einzelnen Schülers bei einer bestimmten Einheit (Einheit x; in Spalte 4 bedeutet 1, dass der Schüler die richtige Antwort wusste, 0 bedeutet eine falsche Antwort). Spalte 5 gibt Noten an auf einer Skala von 1 bis 100. Für diese Aufgabe ist die letzte Spalte irrelevant.

		Ergebnis beim Grammatiktest 1	**Einheit x richtig beantwortet**	**Ergebnis beim Grammatiktest 2**
1	Theo	6	1	55
2	Herbert	5	1	60
3	Martin	8	1	70
4	Kay	7	0	50
5	Vera	7	1	50
6	Lynn	6	0	75
7	Maggie	6	0	60
8	Geoff	8	1	65
9	Rod	2	0	35
10	Petra	3	0	35
11	John	7	1	70
12	Peter	4	0	45
13	Ellis	7	0	75
14	Sara	6	1	60
15	Martin	8	1	90
16	Matty	3	1	60
17	Eve	5	0	60
18	Nancy	5	0	40
19	Adam	8	1	75
20	Mike	7	0	65

   a) Berechnen Sie den *p*-Wert von Einheit x und ihren D-Index.
   b) Wären Sie dafür, die Einheit im Grammatiktest beizubehalten?

2. Nehmen wir an, wir wollen, dass die Verlässlichkeit des Grammatiktests mindestens 0,86 beträgt. Wie könnten wir das erreichen?