Zeitschrift: Bildungsforschung und Bildungspraxis : schweizerische Zeitschrift für Erziehungswissenschaft = Éducation et recherche : revue suisse des sciences de l'éducation = Educazione e ricerca : rivista svizzera di scienze dell'educazione

Herausgeber: Schweizerische Gesellschaft für Bildungsforschung

Band: 20 (1998)

Heft: 2

Artikel: Von der klassischen Testtheorie zur Generalisierbarkeitstheorie : der Beitrag der Varianzanalyse

Autor: Cardinet, Jean

DOI: https://doi.org/10.5169/seals-786241

Von der klassischen Testtheorie zur Generalisierbarkeitstheorie: der Beitrag der Varianzanalyse

Jean Cardinet

Die Messeigenschaften der psychologischen Tests wurden während langer Zeit mit ad hoc Verfahren ohne wirkliche statistische Basis kontrolliert. Cronbach und seinen Mitarbeitern ist es in den 70er Jahren gelungen, diese Verfahren wieder in den Rahmen der Varianzanalyse zu setzen. Die Qualität der Messung zu schätzen besteht im grunde darin, einen Vertrauensbereich um den wahren Wert zu erstellen. Die Grösse des voraussehbaren Fehlers ist von der Stichprobenvarianz abhängig, die durch die Beobachtungsbedingungen verursacht wird: Moment der Prüfung, Auswahl der Fragen, Person des Prüfers, usw. Die Varianzanalyse erlaubt, diese Effekte zu schätzen und ihren kombinierten Einfluss vorherzusagen.
Das Modell des Generalisierbarkeit zeigt, dass die Reliabilität nicht eine Eigenschaft des Tests ist, sondern der «Generalisierung» (von der Stichprobe auf die Bezugspopulationen), die man vornehmen will. «Absolute» Messungen, wie die von Prüfungen, werden so analysierbar gemacht, während die klassische Theorie nur über «relative» Messungen, wie die von Ausleseverfahren, Aussage machen kann.

Die psychologischen Tests, wie viele andere Techniken der angewandten Psychologie, sind Moden unterworfen. Um eine lange Debatte über die Zweckmäßigkeit ihrer Anwendung zu vermeiden, halten wir uns an eine Tatsache: Diese Tests existieren und können in bestimmten Fällen von Nutzen sein. Indem sie generell quantitative Messungen anbieten, haben sie insbesondere zwei Vorteile: Sie können zugleich klare und nuanzierte Angaben liefern, und darüber hinaus kann noch die Genauigkeit ihrer Resultate kontrolliert werden.

Bei richtiger Anwendung dieser Techniken, müssen die Psychologen einen Vertrauensbereich um die beobachteten Werte bestimmen können. Dieser ist natürlich von der Genauigkeit ihrer Meßinstrumente abhängig.

Eine solche Kontrolle wird leider allzuoft vorgenommen, ohne die neuesten theoretischen Entwicklungen zu berücksichtigen, und zwar auf der Basis einer Reliabilitätstheorie, die auf den Anfang dieses Jahrhunderts zurückgeht. Das Festhalten der Psychologen an der Vergangenheit resultiert zweifellos aus der Befürchtung, Berechnungen machen zu müssen, die ihnen als zu komplex erscheinen.

Die aktuelle Entwicklung der Informationstechnologie erlaubt jedoch, dieses Problem zu umgehen, vorausgesetzt man versteht die Logik des angebotenen Modells und stellt dem Computer die richtigen Fragen.

Auf den folgenden Seiten wollen wir zeigen, daß es möglich ist – ganz ohne Formeln – eine mit der modernen Statistik kompatible Testtheorie vorzustellen: die *Generalisierbarkeitstheorie.* Alle Benutzer von Tests oder Prüfungen sollten demnach heute in der Lage sein, die Fehlerspanne ihrer Instrumente richtig zu berechnen.

Wir gehen zuerst auf die Mängel der klassischen Theorie der psychologischen Tests ein und behandeln dann das statistische Modell der Varianzanalyse, das sie ersetzen kann. Wir zeigen, wie das Modell der Generalisierbarkeit deren Probleme umgeht, und wie es erlaubt, darüber hinauszugehen, indem es neue, wichtige Anwendungen ermöglicht, deren soziale Dringlichkeit nicht zu leugnen ist. In einem kurzen Abschluß stellen wir die Generalisierbarkeitstheorie anderen, konkurrierenden Theorien gegenüber.

Die klassische Theorie

Geschichte der psychologischen Tests

Der geniale Einfall Galtons

Galton ermöglichte es, so verschiedene Leistungen wie Schallempfindlichkeit und Klopfgeschwindigkeit zu vergleichen, indem er die Verteilung der Resultate von zahlreichen Versuchspersonen aufstellte, die mit einer ähnlichen Aufgabe konfrontiert waren. Allein die relative Position der Personen in der Verteilung wurde für die Interpretation verwendet.

Konsequenz: feste, situationsgebundene Instrumente

Diese Methode machte eine strenge Standardisierung der Prüfungsbedingungen notwendig. Diese Standardisierung ist für alle quantitativen psychologischen Prüfungen obligatorisch geblieben.

In anderen Situationen, insbesondere bei beruflichen Fachprüfungen, müssen im Gegenteil die Bedingungen genügend variieren können, damit die Beobachtungen einen gewissen Bezug zur betreffenden sozialen Realität behalten, die immer vielfältiger ist, als der im Laufe der Ausbildung behandelte Stoff. Die Prüfungsresultate müssen auf eine ganze Reihe von anderen Situationen generalisierbar sein.

Die Meßinstrumente als situationsspezifisch zu betrachten, stellt eine Begrenzung ihres Anwendungsgebietes dar, hat aber noch andere Nachteile: Es verpflichtet dazu, Verkauf und Einsatz der Tests zu kontrollieren. Diese intentionelle Geheimhaltung trägt zu dem negativen Bild bei, das dem Beruf des Psychologen in der Öffentlichkeit bisweilen anhaftet.

Ihrer Theorie fehlt es an statistischer Basis

Nur beschreibende Statistiken

Die Berechnungen, die für die Kontrolle der Meßeigenschaften von Tests nötig sind, behandeln die Daten immer so, als ob sie eine vollständige Population darstellten.

Die klassischen Formeln sehen keinerlei Inferenz von der beobachteten Stichprobe auf die Bezugspopulation vor.

Konsequenz: Anfällige Äquivalenzhypothesen zwischen Tests

Gulliksen, der die klassische Testtheorie systematisiert hat, hat diese auf das Postulat des Vorhandenseins von Paralleltests, von identischen Mittelwerten und Standardabweichungen und gleichen Interkorrelationen basiert, was ihnen identische wahre Varianzen sowie identische Fehlervarianzen gibt.

Bei tatsächlich erhobenen Daten ist die Äquivalenz der Parallelformen nie perfekt. Verfügt man über kein statistisches Modell, das angibt, was die Zufallsauswahl von Personen- und Fragenstichproben an zufälligen Schwankungen produzieren kann, weiß man nicht, inwieweit man berechtigt ist, die Äquivalenzbedingung als erfüllt zu betrachten.

Vielfalt der ad hoc Koeffizienten

Test-Retest-Reliabilität, Koeffizient von Paralleltests, Split-Half Koeffizient, interne Homogenität, Übereinstimmung zwischen Korrektoren oder Bewertern ...

Die Psychologen, die natürlich den Wert ihrer Prüfungen kontrollieren wollten, haben auf verschiedene Lösungen zurückgegriffen. Da es sich als schwierig herausstellte, den gleichen Test wiederholt ablegen zu lassen, wie es das Postulat der Unveränderlichkeit von Bedingungen verlangt hätte, haben sie versucht, einen zweiten Test parallel zum ersten zu erstellen. Aber weil auch das nicht unproblematisch war, haben sie ihren Test halbiert, um zwei Schätzungen zu erhalten.

Diese Lösung war ebenfalls anfechtbar, wenn beide Testhälften nicht gleichwertig waren. Ein interner Homogenitätskoeffizient war also vorzuziehen. Damit konnte man zwar die Homogenität der Prüfung schätzen, aber nicht mehr die Wiederholbarkeit der Messung ...

Man sieht, daß das Konzept der Reliabilität mehrere Fragen aufwirft, und daß kein Koeffizient alle diese Fragen zufriedenstellend beantwortet. So ist zum Beispiel die Reliabilität zwischen Korrektoren eindeutig anderer Art als die der

vorhergenannten Koeffizienten. Dennoch ist sie an der Replizierbarkeit der Resultate beteiligt.

Konsequenz: Schwierigkeit diese Informationen zu kombinieren

Um all diese Schwierigkeiten noch zu vergrößern, ist die gleichzeitige Berücksichtigung mehrerer Fehlerquellen ein Problem, das die klassische Theorie nicht zu lösen vermag.

Wenn zum Beispiel mehrere Korrektoren dieselben Arbeiten korrigieren, stellen sich zwei Fragen: Wie kann man die berechenbaren Interkorrelationen zwischen ihren Ergebnissen – paarweise genommen – zusammenfassen? (Die Übereinstimungskoeffizienten sind keine Bravais-Pearson Koeffizienten und sind mit ihnen nicht direkt vergleichbar.) Und vor allem: Wie kann man global abschätzen, ob die erhaltene Messung genug abgesichert ist, wenn man alle störenden Einflüsse kumuliert?

Die Nachteile des Referenzindex «r»

Der Korrelationskoeffizient Bravais-Pearson

Die Reliabilitätskoeffizienten können je nach Situation verschieden sein, aber am häufigsten versucht die klassische Theorie die Korrelation Bravais-Pearson zwischen zwei gleichwertigen Schätzungen des gleichen psychologischen Merkmals zu berechnen.

Konsequenz: Nur relative Skalen

Dieser Index arbeitet aber mit standardisierten Ergebnissen, die einen Durchschnitt von null und eine Standardabweichung von eins haben. Der «r» Koeffizient läßt also jeden Durchschnitts- oder Abweichungsunterschied verschwinden, der zwischen den Resultaten eines Tests und denen seiner Parallelform existieren könnte. Mehrere Formeln, die die Homogenität eines Tests messen, wie der Kuder-Richardson 20 (der auf den Kovarianzen zwischen Items beruht), lassen auch die Schwierigkeitsstufe der Items unberücksichtigt. Die Berechnungsformeln der Reliabilität bestätigen also nur, daß die Rangordung der Versuchspersonen sich von Prüfung zu Prüfung nicht verändert, ziehen aber die wirkliche Zahl von richtigen Antworten in jeder Prüfung nicht in Betracht.

Das heißt, daß die klassische Theorie eine rein relative Meßskala voraussetzt, und daß sie nicht dazu geeignet ist, absolute Skalen zu behandeln. (Absolute Skalen sind solche, die in Prüfungen verwendet werden. Für Prüfungen an den Schulen der Provinz Quebec, zum Beispiel, fordert das Gesetz 60% richtige Antworten als Erfolgskriterium. Viele Länder legen auch einen Maximalsatz an Fehlern in den theoretischen Fahrprüfungen fest.) Die klassische Theorie erlaubt es jedoch nicht, die Reliabilität eines Tests unter Berücksichtigung eines solchen absoluten Wertes zu berechnen.

Auswirkung auf die Psychometrie

Möglichkeiten präziser Messungen, die von der Psychometrie in zahlreichen Gebieten angeboten werden

Die klassische Testtheorie hat zweifellos eine äußerst nützliche Rolle gespielt, indem sie die Replizierbarkeit der Rangordung zu kontrollieren ermöglichte, die für die Versuchspersonen erhalten wurden. Sie hat damit wertvolle Qualitätsnormen für die relativen Skalen festgelegt, bleibt aber in einer Wettbewerbslogik eingeschlossen.

Erforderliche Differenzierung der Versuchspersonen

Wie Galton es als erster dargestellt hatte, ist es immer die Streuung der Beobachtungen, die die Interpretation der Ergebnisse begründet. Wenn diese Streuung nicht breit genug ist, funktioniert die Testmethode nicht mehr. Prüfungen in Schulen, zum Beispiel, die auf die Inhalte der Grundlernziele beschränkt wären, würden es nicht erlauben, die Schüler replizierbar einzustufen. Eine Rangordnung kann nur durch ein Überschreiten der Grundlernziele hergestellt werden.

Konsequenz: Konkurrenzlogik, Umwandlung der Prüfungen in Ausleseverfahren

Gewissermaßen fabriziert die Schule den Mißerfolg selbst, indem sie überwiegend Prüfungen einsetzt, die Unterschiede zwischen Kindern schaffen, den Kontrast zwischen starken und schwachen Schülern betonen, und somit die sozio-kulturell begünstigten Schichten bevorzugen.

Die Schule stellt also nicht nur Unterschiede fest, sondern schafft sie mit solchen Prüfungen, die automatisch selektiv sind, da die Qualität der Messung an diese Selektivität gebunden ist.

Einer der Vorzüge der Generalisierbarkeitstheorie besteht in dem Versuch, die Verfasser von pädagogischen Prüfungen aus dieser Wettbewerbslogik herauszuholen.

Die Theorie der Generalisiebarkeit

Ihre Entwicklung

Ausgangspunkt

Im Jahre 1941 liefert Hoyt eine Formel, um die Reliabilität eines Tests mit der Varianzanalyse zu schätzen. Andere Forscher, wie Lindquist 1953, entwickeln seine Arbeiten weiter.

Cronbach, Rajaratnam und Gleser veröffentlichen 1963 eine «Neuinterpretation» der Reliabilitätstheorie. Diese Autoren begreifen eine Reliabilitätsstudie als die Kontrolle des Generalisierbarkeitswertes, den man erhält, wenn man auf

der Basis einer Stichprobe ein Urteil über ein ganzes Universum ähnlicher Beobachtungen fällt.

Es gelingt diesen Autoren, alle vorher existierenden Verfahren als Sonderfälle ihres generellen Modells darzustellen: der «Generalisierbarkeit».

Erweiterungen

In den Jahren 1976, 1981-82 und 1983, veröffentlichen Cardinet, Tourneur und Allal Erweiterungen der Theorie, die die Berechnung der Generalisierbarkeit für beliebige Facetten (nicht nur die der Versuchspersonen) und für beliebige Erhebungsmodi von Stichproben erlauben. Sie nehmen das Konzept der Zuverlässigkeit aus dem Bereich der Psychometrik heraus und machen so die Generalisierbarkeit auf jedes Meßproblem anwendbar.

Cardinet und Tourneur geben 1985 ein Buch heraus, das diese Arbeiten zusammenfaßt: «Assurer la mesure». Die Anwendung der Techniken der Generalisierbarkeit wird leicht verständlich von Bain & Pini in ihrer «Gebrauchsanweisung» von 1996 erklärt.

Ihre statistische Basis: Die Varianzanalyse

Das feste Modell (fixed model)

Kann man die ganze Population beobachten, so stellt die Varianzanalyse kein Problem dar.

Das Modell der ANOVA läßt gelten, daß jedes beobachtete Ergebnis die Summe von positiven und negativen Effekten ist, die den verschiedenen Niveaus der Faktoren des Analyseplans zuzuschreiben sind.

Auf der einfachsten Beobachtungsebene, auf der alle Versuchspersonen die gleichen Items beantworten, geht die Analyse von der Kreuztabelle aus, die die Versuchspersonen in den Zeilen und die Items in den Spalten erscheinen läßt. Jeder beobachtete Wert ist also die Summe des Gesamtmittelwertes plus einem Zeileneffekt, der den Generalfactor der Versuchsperson darstellt, plus einem Spalteneffekt, der den Schwierigkeitsgrad des Items wiedergibt, plus einer Interaktion Zeile/Spalte, die dem Sachverhalt entspricht, daß das entsprechende Item für den entsprechenden Schüler besonders leicht oder besonders schwer sein kann. Die Summe aller dieser Wertekomponenten setzt sich aus den beobachteten Resultaten zusammen.

Nimmt man das Quadrat jedes der erwähnten Wertekomponenten, die einem Effekt entsprechen, sagen wir dem Linieneffekt (Generalfaktor), und errechnet man den Mittelwert, so erhält man die Varianzkomponente für diesen Effekt.

Besteht ein Faktor zum Beispiel aus nur zwei Meßniveaus, wie das Geschlecht, so ist die Varianzkomponente für einen Jungen oder für ein Mädchen das Quadrat des Effekts (positiv oder negativ), das dem Geschlecht zuzuschreiben ist.

Das Zufallsmodell

Kann man nur Stichproben von Versuchspersonen und von Items untersuchen, wird die Sache komplizierter, weil in diesem Falle die Schwankungen der Stichprobenerhebung zu den wahren Varianzen hinzukommen.

Kennt man aber die Stichprobengröße und die Größe der Grundgesamtheit, kann man den Erwartungswert dieser zufallsbedingten Stichprobeneffekte schätzen. Man kann also korrigierte Schätzungen der Varianzkomponenten berechnen, die von diesen Stichprobeneffekten bereinigt sind.

Das generelle Modell von Cornfield und Tukey

Die Generalisierbarkeitstheorie benützt ein statistisches Modell, das für begrenzte wie für unbegrenzte Universen verwendbar ist, und berechnet auf dieser Basis für alle Haupt- und Interaktionseffekte Varianzschätzungen.

Konsequenz: Faktoren, die man kennen sollte

Die Varianzanalyse gibt zuerst über die Existenz und über die Bedeutung der untersuchten Varianzquellen Auskunft. Es kann vorkommen, daß gewisse Effekte fast ohne Auswirkung bleiben, während andere den größten Teil der beobachteten Variationen erklären. Es ist natürlich wichtig, das zu wissen.

Die Varianzanalyse erlaubt auch zu sehen, ob der Stichproben-Erhebungsmodus die Resultate entscheidend beeinflußt hat, denn es ist manchmal möglich, diese Stichprobenerhebung zu ändern, um die Fehler zu reduzieren oder die Varianz, die uns interessiert, zu vergrößern.

Der Beitrag dieser Berechnungen erweist sich als besonders interessant, wenn man versucht, in Hinblick auf seine Anwendung auf konkrete Entscheidungen, den Plan zu optimieren. Das Instrument (oder der Beobachtungsraster), mit dem man arbeiten will, kann in seiner Länge oder in seiner Struktur angepaßt werden, wenn man weiß, woher die ausgeprägtesten Störungen kommen.

Ihr theoretisches Unifizierungsmodell

Die Beobachtungsbedingungen werden als Zufallsstichprobe betrachtet

Anstatt als ein eher mystischer wahrer Wert dazustehen, wird der Universalwert als der Mittelwert aller möglichen Beobachtungen für ein gut abgegrenztes Universum von Beobachtungsbedingungen definiert.

Die Fehler entstehen durch die Zufallsziehung der Facettenniveaus und den daraus resultierenden Interaktionen.

Vom Konzept her ist das Modell hier wieder klar, auch wenn in der Praxis die Beobachtungen nur selten wirklich zufällig gezogen werden. Damit das Modell zufällig wird, genügt es, daß die Beobachtungen austauschbar sind.

Wenn die Wahl der Items oder der anderen Niveaus durch den Zufall bestimmt sind, weiss man, welche Schwankungen zu erwarten sind. Prüfungen

können in diesen bekannten Grenzen parallel sein, ohne im strikten Sinne gleichwertig zu sein.

So kann man aus einer Itembank sukzessive Testformen ziehen, und die Stichprobenheorie kann uns die Erwartungsvarianz der Schwierigkeitsgrade dieser Formen geben.

Die klassischen Reliabilitätskoeffizienten werden auf ein gemeinsames Schema gebracht.

Der Koeffizient «Test-Retest» entspricht dem Effekt der Facette «Zeitpunkte». Der Split-Half Koeffizient entspricht dem Effekt der Facette «Items», etc.

Es ist möglich, die Zeitpunktevarianz der Itemvarianz hinzuzufügen, da in der Statistik die Varianzen unabhängiger Effekte addiert werden.

Mehrere Schlüsse sind auf der Basis eines gegebenen Wertes möglich

Die klassischen Reliabilitätskoeffizienten ließen vielleicht an die Möglichkeit einer einzigen Schätzung der Reliabilität für ein Instrument glauben. Das Konzept der Generalisierbarkeitstheorie, das aus der Statistik hervorgeht, zeigt jedoch, daß eine Beobachtung gleichzeitig verschiedenen Universen angehören kann, wenn diese ineinander nisten. Ein Schüler kann zum Beispiel einer Klasse, einer Schule und einem Bezirk angehören. Sein Ergebnis gibt Auskunft über den Mittelwert dieser drei Universen, jedoch mit mehr Gewicht auf der Schätzung des Klassendurchschnittes als auf der des Bezirks.

Die Reliabilität der Prüfung ist also mehr mit dem Universum verbunden, das man mit der Prüfung schätzen möchte, als mit den Fragen selbst. Wenn man den Mittelwert von Fragen gleicher Kategorie schätzen will, ist die Generalisierbarkeit gut, vorausgesetzt die Schwierigkeit dieser Fragen ist homogen genug. Wenn das Universum verschiedene Kategorien mit sehr verschiedenen Schwierigkeiten umfaßt, ist das Ergebnis nicht zufriedenstellend.

Die Fehlerquellen werden sichtbar

Eine Beobachtung kann nur das sein, was sie ist. Den Fehler zu reduzieren oder zu vergrößern würde bedeuten, eine andere Beobachtung heranzuziehen.

Der Fehler besteht nicht darin, daß der Korrektor sich bei der Berechnung des Totals geirrt hätte. Der Mittelwert des Schülers wird beeinträchtigt, weil die Zufallsziehung zur Folge hat, daß die Beobachtung sich eben auf dieses besondere Resultat konzentriert, bei dem der Korrektor sich geirrt hat.

Die Kontrolle des Fehlers erfolgt nicht durch eine Zweitkorrektur der Arbeit (denn wir wissen nicht, wo er liegt), sondern durch das Heranziehen einer genügenden Anzahl an Arbeiten für jeden Schüler, um die Auswirkung aller zufallsbedingten Störungen zu reduzieren.

Die Vorteile des Referenzindex «Rho Quadrat»

Der Intraklassen-Korrelationskoeffizient

Welche Art Generalisierung auch angestrebt wird, der Koeffizient, der den Wert dieses Urteils mißt, ist immer ein Varianzverhältnis.

Die wahre Varianz wird mit der Varianzkomponente der Versuchsobjekte, meistens die Versuchspersonen, geschätzt. Aus dieser Varianz zwischen Versuchspersonen hat man den Effekt der zufallsbedingten Schwankungen der Beobachtungsbedingungen herausgenommen. Was übrig bleibt, kann als eine Schätzung der Varianz der Universalwerte betrachtet werden.

Die Fehlervarianz ist die Stichprobenvarianz, die durch die Beobachtungsbedingungen entsteht, die in der gewünschten Generalisierung mitberechnet werden (andere Items, andere Momente, andere Korrektoren, etc.), sowie durch die Interaktionen zwischen Facetten.

Die Erwartungsvarianz der beobachteten Werte ist die Summe der beiden vorher genannten Ausdrücke (wahre Varianz plus Fehlervarianz).

Der Koeffizient *Rho Quadrat* ist das Resultat der Teilung der Varianzkomponente der Studienobjekte durch diese Summe. Er stellt einfach einen Prozentsatz der wahren Varianz dar. Er kann also zwischen 0 und 1 variieren.

Cronbach, Rajaratnam und Gleser (1963) beweisen, daß der Intraklassen-Korrelationskoeffizient (der das Verhältnis dieser Varianz zur totalen Varianz schätzt) eine untere Grenze für die Reliabilität des verwendeten Meßinstruments darstellt, egal ob die Fragen die gleichen sind (relatives Rho Quadrat), oder für jede Versuchsperson zufällig gewählt werden (absolutes Rho Quadrat).

Die Schätzung der totalen Varianz, die im Nenner erscheint, wird ausgeführt, indem man der Varianzkomponente «Versuchspersonen» einfach die Stichprobenvarianz hinzufügt, im ersten Fall für relative Messungen, im zweiten für absolute.

Es gibt folglich so viele Rho Quadrate wie es mögliche Generalisierbarkeitsuniversen gibt: eins für jede Stichprobenvarianz, die man in den Nenner stellen kann.

Man kann so tatsächlich auf das Universum wiederholter Messungen generalisieren, entweder anderer möglicher Fragen, oder anderer Korrektoren, oder noch anderer Korrektoren, die mit anderen Fragen prüfen. Die Fehlervarianz ist jedesmal verschieden.

Konsequenz: Andere Reliabilität für relative und absolute Skalen

Im Falle der relativen Meßskala ist die Stichprobenvarianz das Quadrat des Standardmeßfehlers der klassischen Theorie.

Ist die Meßskala aber absolut, sind die Schwankungen in der Schwierigkeit der Fragen eine zusätzliche Fehlerquelle, die der vorigen hinzugefügt wird.

Die Stichprobenvarianz, die die Messungen beeinträchtigt, wird größer. Die Generalisierbarkeit ist also zwangsläufig weniger gut.

... In der Psychologie, um Fehlerbereiche von Prüfungen mit absoluter Skala zu berechnen.

Leistungsmessungen, die eine absolute Skala besitzen, sind in der Psychologie häufig anzufinden: Reaktionszeit, Empfindungsschwelle, Fehleranzahl, etc. Die Stichprobenschwankungen, die solche Messungen beeinträchtigen, sind größer als es der Standardmeßfehler der klassischen Theorie glauben läßt.

Das für die Beobachtung verwendete Instrument kann tatsächlich seinen eigenen systematischen Fehler einbringen, der die Vergleiche zwischen Versuchspersonen nicht beeinträchtigt (da es sich um einen für alle Personen identischen Effekt handelt), sich aber in der absoluten Messung niederschlägt. Die Berechnung eines Fehlerbereiches um einen beobachteten Wert ist das geeignetste Mittel, die Gesamtheit der Stichprobenschwankungen auszudrücken. Um ihn zu erhalten, muß man auf die Generalisierbarkeitstheorie zurückgreifen.

... In der Prüfungslehre, um neue Prüfungsformen zu entwickeln

Es ist klar, daß die Fragen sukzessiver Durchgänge der gleichen Prüfung nicht den gleichen Schwierigkeitsgrad haben können. Die Variation der Prüfungsschwierigkeiten für die Kandidaten eines gleichen Diploms verschiedener Jahrgänge addiert sich also zu dem Meßfehler eines einzigen Prüfungsdurchganges. Aus diesem Grund ist die Generalisierbarkeit einfacher Prüfungen sehr viel kleiner als die von Ausleseprüfungen.

Weil die erdenklichen Fragen zu einem gleichen Thema sehr verschiedene Schwierigkeiten aufweisen können, ist es praktisch unmöglich festzustellen, ob ein Kandidat ein Kompetenzgebiet völlig beherrscht oder nicht. Sein Erfolg ist hauptsächlich von den Fragen abhängig, die ihm gestellt werden.

In seinen Studien zum französischen Baccalaureat hatte Piéron schon gezeigt, daß der Prüfungserfolg mehr vom Prüfer als vom Geprüften abhängt, ohne jedoch die Hauptfehlerquelle in Betracht zu ziehen: die Stichprobenerhebung der Fragen. Die Generalisierbarkeitstheorie erlaubt es, deren Auswirkung zu messen. Man erhält so erschreckende Resultate, daß die Jurys es vorziehen, Prüfungen in Ausleseverfahren umzuwandeln, indem sie im voraus für jeden Durchgang die ungefähre Erfolgsquote festlegen. Das hat zur Folge, daß der Wettbewerb in die Schulen eindringt und das Funktionieren des gesamten pädagogischen Systems negativ beeinflußt.

Es wäre wichtig, diesen Sachverhalt hervorzuheben, indem man um das theoretische Erfolgskriterium Vertrauensbereiche berechnet, um die Hoffnungslosigkeit aller Reformversuche von Prüfungen begreifbar zu machen, an denen erfolglos in allen Ländern gearbeitet wird. Man muß andere Zertifikationsmöglichkeiten suchen, die nicht am Ende wieder zu einer Entscheidung über Bestehen oder Nichtbestehen führen, an der der Zufall den größten Anteil hat.

Man muß sich auf die Beobachtungen der Lehrer stützen. In Großbritannien zum Beispiel bekommen alle Schüler am Ende ihrer Schulzeit eine Bescheinigung, die genau angibt, wozu sie in ungefähr 20 verschiedenen theoretischen und praktischen Bereichen fähig sind. So entstehen nicht mehr zwei Klassen von jungen Leuten (die Diplomierten und die anderen, die zur Arbeitslosigkeit verurteilt sind), sondern alle können bestimmte Kompetenzen vorzeigen, die in verschiedenen Gebieten der Arbeitswelt oder des Studiums geltend gemacht werden können.

... In der Ausbildung, in Situationen, in denen die Varianz zwischen Versuchspersonen gering ist

Es kann vorkommen, daß der Lernprozess so erfolgreich verläuft, daß alle Schüler die geforderte Leistungsschwelle erreichen, oder daß zumindest die Variationen sehr schwach sind, und somit eine Differenzierung zwischen Schülern nicht mehr möglich ist.

In dieser Situation, die wohl das Ideal darstellt, das das öffentliche Erziehungssystem anstreben sollte, scheint die Messung eine ungenügende Reliabilität zu haben, jedenfalls gemäß der klassischen Theorie, die versucht, eine Korrelation zwischen beobachteter und wahrer Rangordnung der Versuchspersonen zu erstellen.

Für die Generalisierbarkeitstheorie dagegen genügt es, einen Fehlerbereich auf der Basis der Stichprobenvarianz der absoluten Messungen zu berechnen, um eine Vorstellung von der tatsächlichen Qualität der Messung zu bekommen.

So wird es endlich möglich, den Erwerb eines Fundamentums bei der großen Mehrheit der Schüler zu kontrollieren, ohne für deren replizierbare Einordnung zusätzliche, extrakurrikulare Schwierigkeiten erfinden zu müssen.

... In der Ausbildung, um Messungen von individuellen Fortschritten zu erlauben

Es gibt Situationen, in denen keinerlei Vergleich zwischen Versuchspersonen erwünscht ist, für die man aber trotzdem die Generalisierbarkeit der Beurteilungen kontrollieren möchte. Das ist der Fall bei Lernkontrollen, bei denen man wissen möchte, ob die Messung des beobachteten Fortschrittes genügend abgesichert ist.

Eine individualisierte Bewertung dieses Typs setzt voraus, daß man die Varianz zwischen Versuchspersonen völlig ignoriert. Dies scheint sowohl für die klassische Theorie als auch für die modernen Meßtheorien, wie die Item Response Theorie, unmöglich zu sein. Die Generalisierbarkeitstheorie dagegen kann dazu eingesetzt werden, die Gesamtheit der Komponenten der Varianz zwischen den Personen und innerhalb der Personen zu errechnen. Auf der Basis dieser Resultate ist es nachher nur noch ein einfaches Rechenproblem, die Reliabilität des beobachteten Lernzuwachses in einem individualisierten Bewertungssystem zu schätzen, in dem jede Versuchsperson nur noch mit sich selbst verglichen wird.

(Man erhält so die durchschnittliche Reliabilität dieser individuellen Messungen. Eine Reliabilitätsberechnung für jede einzelne Versuchsperson wäre möglich, könnte sich aber nicht auf eine ausreichende Anzahl von Daten stützen, um zuverlässige Resultate zu liefern.)

... In der Soziologie, um Meinungen und Kompetenzen zu untersuchen

Die Symmetrie des Modells der Generalisierbarkeit erlaubt hier eher eine Anwendung auf die Differenzierung von Fragen als auf die Differenzierung von Schülern. Man könnte zum Beispiel in einer Befragung kontrollieren wollen, welche Meinungen weit verbreitet sind, und welche Minderheitsstandpunkte darstellen. Was das Funktionieren des Schulsystems betrifft, könnte man auch mittels Surveys kontrollieren wollen, welche Lernziele von der Gesamtheit der Schüler erreicht wurden, und welche Kenntnisse noch Schwierigkeiten bereiten. In allen Arten von Befragungen (egal ob sie Meinungen oder Leistungen betreffen), liefert die Berechnung von Fehlerbereichen um die Mittelwerte bestimmter Fragen eine Basis für die Schätzung der «Spanne», ohne die eine angemessene Interpretation nicht möglich ist.

... In der Epistemologie, um Stadien beim Lernen eines logischen Prinzips zu differenzieren

Für die Forschung in der genetischen Psychologie ist es wichtig, sukzessive Etappen beim Lernen von bestimmten Begriffen oder bestimmten Prinzipien differenzieren zu können. Auch hier sind Fehlerbereiche das geeignetste Mittel, den Wert und die Wiederholbarkeit der festgestellten Lernstufen zu kontrollieren.

... In der Biologie, für die Analyse von Resultaten bei experimentellen Versuchsanordungen

Die Varianzanalyse wird oft herangezogen, um Resultate von Experimenten zu interpretieren. Wendet man die Analyse der Generalisierbarkeit auf gerade diese Daten an, liefert sie einen Fehlerbereich, der den beobachteten Effekten entspricht. Man erhält so ein nuanzierteres Resultat, als wenn man nur das Fehlerrisiko feststellen würde, das man eingeht, wenn man die Nullhypothese akzeptiert.

Man kann sich ohne Schwierigkeiten andere Anwendungsgebiete der Generalisierbarkeitstheorie vorstellen: In der Wirtschaft für Marktstudien, in der Medizin zur Kontrolle von Behandlungseffekten, in der Ethologie für Studien zur Allgemeingültigkeit gewisser Verhaltensmuster, etc., immer dann, wenn Ergebnisse

von Stichproben verwendet werden, um Urteile über die Grundgesamtheit zu fällen.

Die Besonderheit der Neuformulierung, die von Cronbach ausging, zeigt sich vor allem in der Möglichkeit, Versuchspersonen nach anderen Gesichtspunkten als dem Generalfaktor zu differenzieren und damit eine individualisierte Evaluation des Lernvorgangs zu begründen. Darin besteht einer der wenigen Wege, die man dem Schulsystem als Hilfe anbieten kann, sich von dem Zwang zu befreien, unter dem es heute steht, die bestehenden Sozialstrukturen fortzusetzen.

Vergleich der deterministischen Modelle mit den stochastischen Modellen

Die Generalisierbarkeitstheorie gehört der Klasse der deterministischenen Modelle an, in dem Sinne, daß nach einer zufallsbedingten Ziehung ein einziger Wert gefunden wird.

Für die stochastischen Modelle hingegen sind, von einem gegebenen wahren Wert ausgehend, eine ganze Reihe Beobachtungsresultate möglich, von denen man nur die Verteilung kennt, «charakteristische Kurve» des Items genannt. Die Item Response Theorie zieht alle möglichen Schlüsse aus dieser Kurve.

Der Psychologe muß notwendigerweise zwischen diesen theoretischen Ansätzen wählen, denn beide Modelle werden wahrscheinlich noch viele Jahre unvereinbar bleiben. Es wäre deswegen unzureichend, das erste vorzustellen, ohne es kurz mit dem zweiten zu vergleichen. Die Titel des folgenden Paragraphen beziehen sich jedesmal auf die Generalisierbarkeitstheorie.

Die Parameter der Items sind von der Population der Versuchspersonen abhängig und umgekehrt

Der Gesamtmittelwert ist der gleiche für alle Facetten. (Es ist nur möglich, seine Stichprobenvarianz zu schätzen.) Die Resultate der Versuchspersonen oder der Items werden also nicht unabhängig voneinander geschätzt, wie es mit der Item Response Theorie möglich ist.

Die Stichprobenvarianz ist die gleiche für die gesamte Resultatenskala

Die Varianzanalyse setzt voraus, daß die Fehler in allen Zellen des Plans die gleiche Varianz haben. Mit einem Modell, das auf der Varianzanalyse beruht, ist es demnach unmöglich zu behaupten, daß ein gegebenes Item niedrigere oder höhere Kompetenzniveaus präziser mißt. Hier liegt ein weiterer Vorteil der Item Response Theorie gegenüber der Theorie, die wir oben vorgestellt haben.

Aber das Meßsystem kann analysiert und aufgrund einer G-Studie verbessert werden

Die Item Response Theorie erlaubt weder die Quellen der Fehlervarianz zu bestimmen, noch sie zu kontrollieren. Es ist der Hauptvorteil der Generalisierbarkeitstheorie die Möglichkeit einer Optimierung des Plans anzubieten, im Anschluß an eine Voruntersuchung, bei der die beteiligten Varianzkomponenten geschätzt werden konnten.

Das Meßsystem kann für die Bewertung der Leistungen von Klassen, Schulen, Organisationen, etc. angepaßt werden.

Der vielversprechendste Anwendungsbereich für die Generalisierbarkeitstheorie ist zur Zeit das Erstellen von Fehlerbereichen für die Messung von Leistungen in Institutionen, die einer externen Evaluation unterzogen werden.

In den USA werden oft Entscheidungen auf der Basis von Resultaten dieser Evaluationen getroffen. Diese Entscheidungen haben oft schwerwiegende Folgen für die Mitarbeiter dieser Institutionen, ohne daß der Wert der so verwendeten durchschnittlichen Leistungen wirklich abgesichert wurde, wie es Cronbach und seine Kollegen in einem kürzlich erschienenen Artikel (1995) unterstreichen.

Es wäre heute Rolle der Forscher, diese Situation näher zu betrachten, die Fehlerbereiche zu berechnen und die Betroffenen, die Behörden und die Öffentlichkeit zu warnen, wenn die verwendeten Messungen ungenügend abgesichert sind.

Die Ärzte haben ein deontologisches Prinzip: «keinen Schaden zufügen», dem sie seit Hippocrates in allen Situationen zu folgen versuchen. Die Sozialwissenschaftler müssen die gleiche Sorge tragen, was die Konsequenzen der Messungen betrifft, die sie den Verantwortlichen und der Öffentlichkeit überliefern. Wir hoffen, daß die Generalisierbarkeitstheorie dazu beitragen kann, indem sie die Aufmerksamkeit der Forscher auf die Evaluationsverfahren lenkt, deren Zuverlässigkeit nicht bestätigt ist.

Technische Notiz

Das größte Hindernis für die Anwendung der Generalisierbarkeitstheorie war lange Zeit das Fehlen eines Programms, das auf Microcomputern eingesetzt werden konnte. Brennan hat zwar sein Programm GENOVA (ursprünglich für große Rechner geschrieben) für PC angepaßt, dessen Logik aber nicht verändert. Diese sieht eine «Batch»-Arbeit vor. Ferner wurde sein Programm einzig für die Studie der Zuverlässigkeit von Meßinstrumenten konzipiert, die auf Versuchspersonen angewendet werden.

Aus diesem Grunde wurden kürzlich zwei andere Programme entwickelt, deren Anwendung allgemeiner ist: Eins für Macintosh und eins für PC. Das erste wurde an der Universität Laval, Provinz Quebec (Alain McNicoll et Françoise Cordeau CESSUL, 1996), das zweite mit der Unterstützung der Universitäten Freiburg und Genf (Pierre Ysewijn, 1996) erstellt.

Von beiden Programmen (kostenlos erhältlich) existiert eine französische und eine englische Fassung. Die Programme für Macintosh (Etudgen und Gen) sind über die Internetadresse (http://www.cessul.ulaval.ca) zugänglich. Die Programme für DOS (GT 2.0.F und GT 2.0.E) stehen auf der Seite des Institut Romand de Recherches et de Documentation Pédagogiques (IRDP) in Neuchâtel, Internet-Adresse: (http://www.unine.ch/irdp/stat/generali.htm) zur Verfügung.

In einer Fremdsprache arbeiten zu müssen, kann noch ein Handikap für die Anwendung des Modells der Generalisierbarkeit in deutschsprachigen Ländern sein. Aus diesem Grund hat *die Stiftung Suzanne und Hans Biäsch zur Förderung der Angewandten Psychologie* die deutsche Übersetzung des PC-Programms (GT 2.0.D) subventioniert. Die deutsche Version, die auf DOS arbeitet, kann, wie die französische und die englische, von der Internet Seite des IRDP heruntergeladen werden.

Die Zip-gedrängt Datei enthält ein Handbuch, in MS-Word Form und 12 Übungsdateien.

Bibliographie

Bain, D., & Pini, G. (1996). *Pour évaluer vos évaluations - La généralisabilité: mode d'emploi.* Genève: Centre de Recherches Psychopédagogiques.

Brennan, R. L. (1992). *Elements of generalizability theory* (2nd ed.). Iowa City (IA): ACT Publications.

Cardinet, J., & Allall, L. (1983). Estimation of generalizability parameters. In L. J. Fyans (Eds.), *Generalizability theory: Inferences and practical applications* (pp. 17-48). San Francisco: Jossey-Bass.

Cardinet, J., & Tourneur, Y. (1985). *Assurer la mesure.* Berne: Peter Lang.

Cardinet, J., Yvan, T., & Allal, L. (1976). The symmetry of generalizability theory: applications to educational measurement. *Journal of Educational Measurement,* 13 (2), 119-135.

Cardinet, J., Tourneur, Y., & Allal, L. (1981). Extension of generalizability theory and its applications in educational measurement. *Journal of Educational Measurement,* 18(4), 183-204; and 1982, 19, p. 331-332.

Cornfield, J., & Tukey, J. W. (1956). Average values of mean squares in factorials. *The Annals of Mathematical Statistics,* 27 (4), 907-949.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. (1995). *Generalizability analysis for educational assessments.* Evaluation Comment, Summer issue, 1-29. (Available electronically from http://www.cse.ucla.edu.)

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: a liberalization of reliability theory. *British Journal of Statistical Psychsychology,* 16, 137-163.

Gulliksen, H. (1950). *Theory of Mental Tests.* New York: John Wiley & Sons.

Hoyt, C. (1941). Test reliability estimated by analysis of variance. *Psychometrika,* 6, 153-160.

Lindquist, E. F. (1953). *Design and analysis of experiments in psychology and education.* Boston: Houghton-Mifflin.

McNicoll, A., & Cordeau, F. (1996). ETUDGEN: *Logiciel de généralisabilité- Guide d'utilisation.* Forme I, Québec: Université Laval - Centre d'évaluation des Sciences de la Santé, 28 p.

Piéron, H. (1969). *Examens et docimologie.* Paris: PUF.

Ysewijn, P. (1996). GT: *Logiciel pour études de généralisabilité.* Forme H, CH-1038 Bercher: Pierre Ysewijn.

De la théorie classique des tests à la théorie de la généralisabilité: l'apport de l'analyse de la variance

Résumé

Les propriétés métriques des tests psychologiques ont été contrôlées longtemps par des procédures ad hoc, constituant la théorie classique des tests, sans fondement statistique véritable. Cronbach et ses collaborateurs, dans les années 70, ont réussi à replacer ces procédures dans le cadre de l'analyse de la variance. Evaluer la qualité de la mesure revient en effet à estimer un intervalle de confiance autour du score vrai que l'on cherche à estimer. L'importance de l'erreur à prévoir est fonction de la variance d'échantillonnage causée par les conditions d'observation: moment de l'examen, choix des questions, personnalité de l'observateur, etc. L'analyse de la variance permet d'estimer ces effets et de prévoir leur influence combinée.

Selon la théorie de la généralisabilité, la fidélité d'un examen n'est pas une propriété du test lui-même, mais une fonction de la «généralisation» que l'on veut faire (de l'échantillon aux diverses populations parentes que l'on peut envisager). La fidélité de mesures «absolues», comme celles des examens, peut alors être contrôlée, alors que la théorie classique des tests ne pouvait traiter que de mesures relatives, comme celles des concours.

Della teoria classica dei test alla teoria della generalizzabilità : il contributo dell'analisi della varianza

Riassunto

Per molto tempo, la qualità metriche di un test psicologico sono state controllate utilizzando procedure ad hoc, che appartengono alla teoria classica dei test ma che non hanno un vero fondamento statistico. Durante gli anni 70, Cronbach ed i suoi collaboratori sono pervenuti a sostituire queste procedure partendo dai principi che reggono l'analisi della varianza. Valutare le qualità della misura significa infatti stabilire un intervallo di confidenza (intervalle de confiance, confidence interval) attorno al risultato vero che si vuole stimare. L'importanza dell'errore prevedibile associato a questa stima dipende dalla varianza dovuta alla scelta aleatoria delle condizioni di osservazione : Momento dell'esame, scelta delle domande, personalità dell'esaminatore, ecc. L'analisi della varianza permette si stimare questi effetti e di prevedere l'entità della loro influenza combinata.

Secondo la teoria della generalizzablità, la «fedeltà» (o precisione: fidelité, reliability) di un esame non è una proprietà del test in quanto tale, ma piuttosto una funzione della «generalizzazione» che si vuol fare (dal campione a diverse popolazioni che possono essere considerate). La precisione di misure «assolute» (comme quelle fornite da un esame) può così essere controllata, contrariamente a quanto accade nell'ambito della teoria classica dei test che analizza unicamente misure relative (comme quelle che riguardano un concorso per esempio).

From classical test theory to generalizability theory: the contribution of ANOVA

Summary

For many years, the metric properties of psychological tests were controlled through ad hoc procedures, lacking proper statistical basis. In the seventies however, Cronbach and his coworkers succeded in relating reliability formulas to the theoretical framework of ANOVA. Estimating the quality of a mesasuring instrument involves setting the limits of a confidence interval around the true score. The magnitude of the probable error depends on the sample variance resulting from the random choice regarding the conditions of observation:

moment chosen for testing, selection of questions, personality of the examiner, etc. The algorithms of ANOVA yield estimates of these effects and enable statisticians to predict the magnitude of their combined influence.

According to the model of Generalizability Theory, the reliability of a testing procedure is not a property of the test itself, but a function of the generalizations that one wants to make (from the sample to various populations of reference). The reliability of «absolute» measurements, typical of scholastic examinations, can now be controlled, while classical test theory could only assess «relative» measurements, usually found in situations of selection.