Diskussionsbeiträge des Fachbereichs Wirtschaftswissenschaft
der Freien Universität Berlin

Betriebswirtschaftliche Reihe

2007/4

Methodische Grundlagen und Anwendungen der Generalisierbarkeitstheorie in der betriebswirtschaftlichen Forschung

Martin Eisend

3-938369-52-3

Zusammenfassung

Die Generalisierbarkeitstheorie stellt einen messtheoretischen Ansatz dar, der im Gegensatz zur klassischen Testtheorie verschiedene Fehlerquellen gleichzeitig berücksichtigen kann. Auf der Basis der Ergebnisse einer Generalisierbarkeitsstudie lassen sich genaue Angaben zur Gestaltung generalisierbarer Messdesigns ermitteln. Der Beitrag gibt eine kurze Einführung in die Generalisierbarkeitstheorie mit ihren Grundannahmen und ihrer typischen Untersuchungsanlage. Darauf aufbauend wird ein Überblick über Anwendungen der Generalisierbarkeitstheorie in der betriebswirtschaftlichen Forschung gegeben.

Summary

Generalizability theory provides an alternative and more flexible approach for construct measurement than classical measurement theory. The theory takes into account several sources of error. Generalizability studies lead to results that allow for designing measurements that optimize the generalizability of results. The paper gives a short introduction into generalizability theory and summarizes applications of generalizability theory in business research.

1 Einleitung

Zur Erfassung komplexer Phänomene werden Messskalen verwendet, die häufig mehrere Items, manchmal auch mehrere Dimensionen umfassen. Die Güte dieser Messskalen wird dabei in der Regel über gängige Kriterien der Validität und Reliabilität erfasst. Ein Aspekt, der dabei meist vernachlässigt wird, ist die Frage nach der Generalisierbarkeit der Messung über verschiedene Bedingungen und Kontexte hinweg. Die Generalisierbarkeitstheorie stellt einen messtheoretischen Ansatz dar, der sich genau dieser Problematik widmet. Nachfolgend wird diese Theorie mit ihren Grundannahmen und ihrer typischen Untersuchungsanlage dargestellt und diskutiert. Darauf aufbauend wird ein Überblick über Anwendungen der Generalisierbarkeitstheorie in der betriebswirtschaftlichen Forschung gegeben. Ziel des Beitrags ist es, eine verständliche Einführung in die Generalisierbarkeitstheorie zu geben und die Breite der Anwendungsmöglichkeiten in der betriebswirtschaftlichen Forschung aufzuzeigen.

2 Probleme der klassischen Testtheorie

Die Generalisierbarkeitstheorie knüpft an der klassischen Testtheorie und den dabei verwendeten Reliabilitätstests an. Zur Bestimmung der Reliabilität wird ein beobachteter Wert (bzw. seine Varianz) in die Komponenten "wahrer Wert" und "Fehler" zerlegt. Je kleiner der Fehler, desto zuverlässiger ist die Messung. Die zur Überprüfung der Reliabilität verwendeten Tests der klassischen Testtheorie gehen jeweils von genau einer bestimmten Fehlerquelle aus. Die verschiedenen Tests aber konzentrieren sich jeweils auf unterschiedliche Fehlerquellen, wodurch streng genommen die in verschiedenen Testverfahren ermittelten Reliabilitätskoeffizienten gar nicht vergleichbar sind (Gleser, Cronbach & Rajaratnam, 1965, S. 396). Folgende Fehlerquellen werden bei den entsprechenden Tests beachtet:

- Test-Retest-Reliabilität → Fehlerquelle Zeit
- Paralleltest-Reliabilität → Fehlerquelle Probanden
- Interne Konsistenzprüfung → Fehlerquelle Items

- Intercoder-Reliabilität → Fehlerquelle Kodierer/Beobachter

Je geringer der jeweilige Messfehler, desto besser lässt sich die Messung über die entsprechende Bedingung verallgemeinern. Insofern machen Reliabilitätstests also auch eine Aussage über die Verallgemeinerbarkeit einer Messung. Messfehler können aber auf verschiedenen Fehlerquellen beruhen, wobei sich diese verschiedenen Fehlerquellen auch gegenseitig beeinflussen können. Dieser Problematik tragen herkömmliche Reliabilitätstests keine Rechnung.

Die genannten Fehlerquellen, die im Rahmen von Reliabilitätstests untersucht werden, sind gekennzeichnet durch die ursprüngliche Entwicklung der klassischen Testtheorie vor dem Hintergrund von Persönlichkeits- und Intelligenztests, wo es um die Messung von Unterschieden zwischen Personen geht. Die Testtheorie stößt an ihre Grenzen bei anderen Untersuchungsobjekten, z.B. wenn einzelne Produkte oder Dienstleistungen als Untersuchungsgegenstand erfasst werden sollen, bei denen ganz andere Fehlerquellen auftauchen können, etwa wenn die Qualität einer Dienstleistung in unterschiedlichen Filialen gemessen wird und diese Filialen somit zur Fehlerquelle der Messung werden können (Finn & Kayande, 1997; Rentz, 1987).

Eine Berücksichtigung verschiedenartiger Untersuchungsobjekte und verschiedener Messfehlerquellen ermöglicht die Generalisierbarkeitstheorie, die von Cronbach und seinen Mitarbeitern entwickelt wurde (Cronbach, Gleser, Nanda & Rajaratnam, 1972; Cronbach, Rajaratnam & Gleser, 1963; Gleser et al., 1965). Diese Theorie verzichtet auf die der klassischen Testtheorie eigenen Äquivalenzannahmen so genannter paralleler Tests, also der Annahme, dass Messungen gleiche Inhalte wiedergeben und gleiche Mittelwerte, Varianzen und Kovarianzen besitzen (Rajaratnam, Cronbach & Gleser, 1965). Diese Annahmen sind bei der der praktischen Umsetzung von Messungen oftmals nicht erfüllt, etwa wenn mehrere Beobachter von Einkaufsverhalten Beurteilungen abgeben, die sich hinsichtlich dessen, worauf sie achten, oder auch im Hinblick auf die Varianz ihrer Beurteilungen durchaus unterscheiden können.

3 Grundlagen der Generalisierbarkeitstheorie

3.1 Statistisches Grundmodell

Grundlegend für die Generalisierbarkeitstheorie ist die Annahme, dass jeder Beobachtungswert eines Untersuchungsobjekts *x* (z.B. die Prüfungsleistung eines Studierenden), eine Stichprobe aus einem Universum möglicher Beobachtungen unter verschiedenen Bedingungen *y, z,…*(z.B. Prüfungszeitpunkt, Prüfungsform) darstellt. Der Erwartungswert bezüglich eines Untersuchungsobjekts über all diese Beobachtungen wird als 'universe score' oder globaler wahrer Wert (μ_x) bezeichnet, was dem wahren Wert in der klassischen Testtheorie entspricht (im Beispiel also die wahre Leistung des Studierenden). Das Universum, auf das die Messung generalisiert werden soll, wird vom Forscher aufgrund theoretischer Vorüberlegungen durch die ihm wichtig erscheinenden Merkmale der Generalisierung festgelegt. Soll also die Leistung von Studierenden (*p*) über verschiedene Prüfungszeitpunkte (*i*) und verschiedene Prüfungsformen (*j*) generalisiert werden, so ergibt sich folgender 'universe score' eines Studierenden:

(1) $$\mu_p = \underset{i,j}{E} X_{pij}$$

Entsprechend der klassischen Testtheorie setzt sich ein beobachteter Wert dann aus diesem 'universe score' und einem Fehlerterm zusammen:

(2) $$X_{pij} = \mu_p + \Delta_{pij}$$

Der Fehlerterm lässt sich nun gemäß der Generalisierbarkeitstheorie im Gegensatz zur klassischen Testtheorie in mehrere Komponenten zerlegen. Dazu bedient man sich der Methode der Varianzanalyse, mit der ein beobachteter Wert in verschiedene varianzanalytische Komponenten zerlegt werden kann, die auf die Effekte unabhängiger Variablen, deren Interaktionen sowie einen Fehlerterm zurückzuführen sind. Entsprechend kann ein Messwert im genannten Beispiel auf die Effekte der Studierenden, der Prüfungszeitpunkte, der Prüfungsformen und deren Interaktionen, anderer systematischer Fehlerterme und der zufälligen Fehlervarianz zurückgeführt und entsprechend zerlegt werden. Für einen Beobachtungswert ergibt sich somit die folgende Strukturgleichung:

(3)

$X_{pij} =$	(a) Beobachtungswert
$= \mu$	(b) Gesamtmittelwert (Erwartungswert über alle Studierenden, Prüfungszeitpunkte und –formen)
$+ \mu_p - \mu$	(c) Effekt des Studierenden
$+ \mu_i - \mu$	(d) Effekt des Prüfungszeitpunkts
$+ \mu_j - \mu$	(e) Effekt der Prüfungsform
$+ \mu_{pi} - \mu_p - \mu_i + \mu$	(f) Effekt der Interaktion Studierender – Prüfungszeitpunkt
$+ \mu_{pj} - \mu_p - \mu_j + \mu$	(g) Effekt der Interaktion Studierender – Prüfungsform
$+ \mu_{ij} - \mu_i - \mu_j + \mu$	(h) Effekt der Interaktion Prüfungszeitpunkt – Prüfungsform
$+ X_{pij} - \mu_{pi} - \mu_{pj} - \mu_{ij} + \mu_p + \mu_i + \mu_j - \mu$	(i) Effekt der Dreifachinteraktion Studierender – Prüfungszeitpunkt – Prüfungsform, konfundiert mit dem Messfehler

Der beobachtete Wert (a) setzt sich also zusammen aus dem

(b) Gesamtmittelwert μ, der eine Konstante darstellt (also die durchschnittliche Leistung aller Studierender über alle Prüfungsformen und Prüfungszeitpunkte),

(c) der Abweichung des Mittelwerts eines Studierenden von diesem Gesamtmittelwert (wenn ein Studierender tendenziell bessere oder schlechtere Leistung erbringt),

(d) der Abweichung des Mittelwerts des Prüfungszeitpunkts vom Gesamtmittelwert (etwa wenn morgens tendenziell bessere Prüfungsergebnisse erzielt werden als spätabends),

(e) der Abweichung des Mittelwerts der Prüfungsform vom Gesamtmittelwert (etwa wenn bei Klausuren tendenziell schlechtere Ergebnisse, bei mündlichen Prüfungen tendenziell bessere Ergebnisse erzielt werden),

(f) der Interaktion zwischen Studierenden und Prüfungszeitpunkt (wenn die Varianz der Leistung der Studierenden auch über den Zeitpunkt variiert),

(g) der Interaktion zwischen Studierenden und Prüfungsform (wenn die Varianz der Leistung der Studierenden auch über die Prüfungsform variiert),

(h) der Interaktion zwischen Prüfungszeitpunkt und Prüfungsform (wenn beispielsweise Klausuren morgens zu besseren Ergebnissen führen als abends, während mündliche Prüfungen bessere Ergebnisse erzielen am Abend als am Morgen),

(i) sowie der Dreifachinteraktion, die konfundiert ist mit dem Messfehler, der weitere systematische Fehlereinflüsse sowie zufällige Fehler umfasst.

Dreifachinteraktion und Messfehler lassen sich nicht trennen, da man das Untersuchungsobjekt (die Studierenden) als eigenen Faktor in die Varianzanalyse mit einbezieht. Für jede mögliche Faktorkombination liegt also nur eine Beobachtung vor, wodurch eine Mittelwert- und Varianzberechnung nicht mehr möglich ist, die nötig wäre, um den Messfehler gesondert zu ermitteln.

Alle Effekte in der Strukturgleichung haben einen Erwartungswert von Null und somit wird auch in der Verteilung der zugehörigen Werte der Mittelwert gleich Null gesetzt. Die Varianz eines Effekts, z.B. die Varianz des Effekts der Studierenden, also die durchschnittliche quadrierte Abweichung der 'universe scores' der Studierenden vom Gesamtmittelwert lautet dann:

$$\sigma_p^2 = E_p(\mu_p - \mu)^2 \tag{4}$$

Die Varianz aller Beobachtungswerte (über Studierende, Prüfungszeitpunkte und Prüfungsformen) lautet dann:

$$\sigma^2(X_{pij}) = \sigma_p^2 + \sigma_i^2 + \sigma_j^2 + \sigma_{pi}^2 + \sigma_{pj}^2 + \sigma_{ij}^2 + \sigma_{pij,e}^2 \tag{5}$$

Auch hier ist das Residuum aufgrund des Studiendesigns vermischt: Der Term beinhaltet den Interaktionseffekt (die Varianz der Leistungen der Studierenden variieren über die Prüfungsformen und Prüfungszeitpunkte) und den Messfehler. Abbildung 1

ist eine Darstellung eines Venn-Diagramms, in dem die Systematisierung der einzelnen Varianzquellen nochmals verdeutlicht wird.

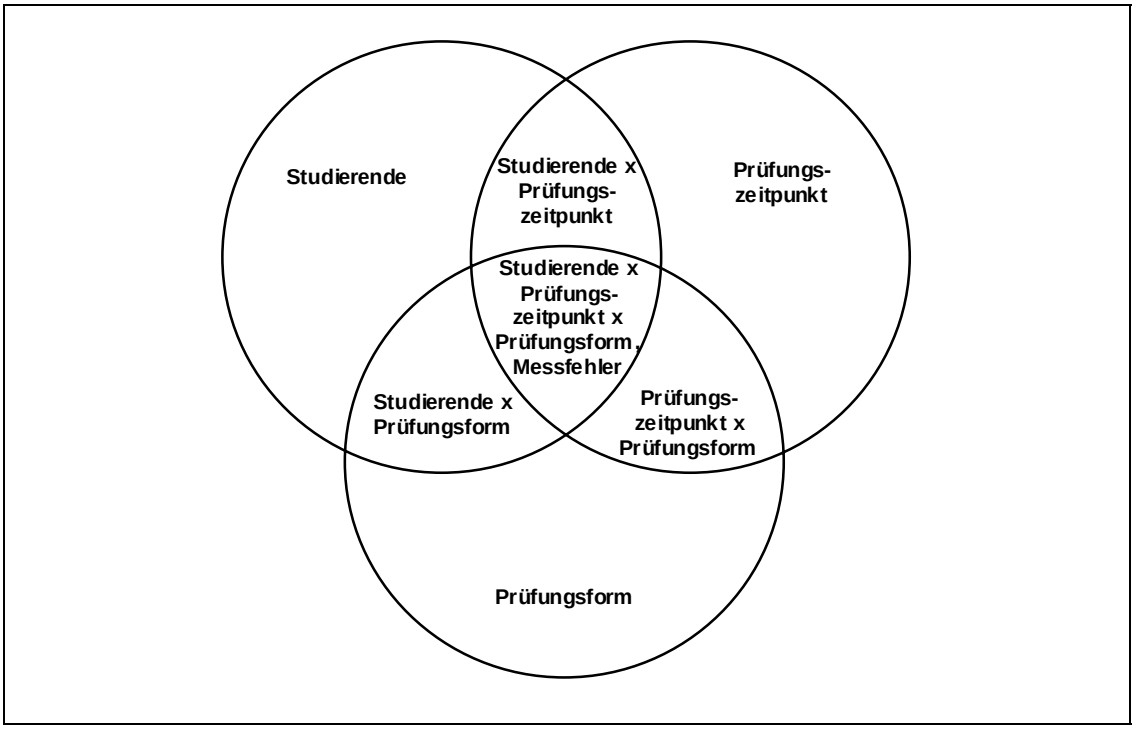

Abbildung 1: Venn-Diagramm für verschiedene Varianzquellen (Studierende, Prüfungszeitpunkt, Prüfungsform) bei der Leistungsmessung

Die Generalisierbarkeitstheorie versucht nun die einzelnen Varianzkomponenten und deren Gewicht im Rahmen einer Generalisierbarkeitsstudie empirisch zu bestimmen. Bei der Anwendung der Generalisierbarkeitstheorie können aus praktischen Gründen meist nur eine begrenzte Anzahl an Fehlerquellen untersucht werden. Stellt diese Teilmenge eine Zufallsauswahl aus allen möglichen Fehlerquellen dar, was als zentrale Annahme der Generalisierbarkeitstheorie gilt (Brennan, 1983, S. 120), so kann man den Erwartungswert der Korrelation von zwei Beobachtungswerten der Korrelation zwischen dem beobachteten Wert und dem 'universe score' gleichsetzen (Campbell, 1976). Wie gut dabei der 'universe score' auf der Basis der vorhandenen Beobachtungen abgebildet wird, wird durch einen Generalisierbarkeitskoeffizienten zum Ausdruck gebracht.

3.2 Untersuchungsanlage

Die Varianzkomponenten der Messung sowie die Generalisierbarkeitskoeffizienten werden empirisch ermittelt. Typischerweise werden bei der Untersuchungsanlage zwei Schritte unterschieden: Die G-Studie ('generalizability study') und die D-Studie ('decision study'). Die G-Studie dient insbesondere der Bestimmung der relevanten Fehlerquellen und deren Beitrag zum Messfehler. Dem Beitrag der Messfehler wird in der D-Studie dann Rechnung getragen. Die D-Studie wird so angelegt, dass die Messfehler reduziert werden und die Messung somit an Generalisierung gewinnt.

3.2.1 Bestimmung der Varianzkomponenten in der G-Studie

Zunächst werden das zu untersuchende Konstrukt und dessen erwünschter Gültigkeitsbereich festgelegt. Im genannten Beispiel soll also der Gültigkeitsbereich der Messung von Prüfungsleistungen von Studierenden möglichst verschiedene Prüfungszeitpunkte und Prüfungsformen umfassen. Dazu bedient man sich eines faktoriellen Designs, wie es im Rahmen der Varianzanalyse eingesetzt wird. Anstatt von Faktoren sprechen Cronbach et al. (1963) von Facetten ('facets') und die einzelnen Faktorstufen werden als Bedingungen ('conditions') bezeichnet. Facetten stellen Merkmale dar, von denen angenommen wird, dass sie zur Varianz der Messwerte beitragen, z.B. die Prüfungsformen, mit denen die Leistung von Studierenden gemessen wird. Die einzelne Prüfungsform (z.B. Klausur, mündliche Prüfung) stellt dabei eine Bedingung dar.

Man unterscheidet weiterhin zwischen Facetten der Differenzierung ('facet of differentiation') und der Generalisierung ('facet of generalization') (Cardinet, Tourneur & Allal, 1976). Die Facette der Differenzierung bezieht sich auf das eigentliche Untersuchungsobjekt, im Beispiel also die Studierenden, während die Facetten der Generalisierung die Fehlerquellen der Messung darstellen. Ein Ziel einer generalisierbaren Messung besteht nun darin, die Varianz der Facette der Differenzierung möglichst groß werden zu lassen im Verhältnis zur Varianz der Facetten der Generalisierung (Hughes & Garrett, 1988): die Leistungen der Studierenden dürfen durchaus variieren, sie sollten aber nicht vom Prüfungszeitpunkt und von der Prüfungsform abhängen. Die Festlegung der Facetten geschieht je nach Untersuchungszweck und kann

auch variieren (Cardinet, Tourneur & Allal, 1981; Cardinet et al., 1976). Misst man beispielsweise die Markentreue von Kunden, so können die Kunden die Facette der Differenzierung bilden, während verschiedene Marken dann zur Facette der Generalisierung gehören. Es können aber auch die einzelnen Marken als Facette der Differenzierung verstanden werden, um somit etwa verschiedene Markenkonzepte und die damit jeweils verbundene Markentreue der Konsumenten unterscheiden zu können (Rentz, 1987).

Beruhen die einzelnen Bedingungen einer Facette auf einer Zufallsstichprobe von möglichen Bedingungen, spricht man von zufälligen Facetten ('random facets'). Werden dagegen alle möglichen Bedingungen einer Facette in einer G-Studie realisiert spricht man von festen Facetten ('fixed'). Die Generalisierbarkeitstheorie geht von zufälligen Facetten aus, die Einbeziehung fester Facetten stellt einen Sonderfall dar, da hier ja eine über die einbezogenen Bedingungen hinausgehende Generalisierung gar nicht mehr erforderlich ist. Daher lässt das Studiendesign der Generalisierbarkeitsstudie auch nicht zu, dass alle Facetten fest sind. Bei der Bestimmung der Varianzkomponenten einer einzelnen festen Facette erfolgt eine Analyse entweder durch eine Mittelwertbildung über alle Bedingungen hinweg, wodurch die feste Facette dann nur mit einer Bedingung repräsentiert wird, oder durch die Durchführung separater G-Studien für jede Bedingung der festen Facetten (vgl. Shavelson & Webb, 1991, S. 66ff.).

Das Design der G-Studie kann vollständig ('crossed') oder unvollständig ('nested') sein, abhängig davon, ob für jede denkbare Kombination von Bedingungen aus den verschiedenen Facetten eine Beobachtung vorgesehen ist oder nicht. So kann bei der Messung der Prüfungsleistung bei jedem Studierenden zu allen Prüfungszeitpunkten mit allen Prüfungsformen gemessen werden (vollständiger Plan) oder aber je Prüfungszeitpunkt wird nur jeweils eine bestimmte Prüfungsform erhoben, z.B. morgens Klausuren und abends mündliche Prüfungen (unvollständiger Plan). Die Facette Prüfungsform ist wäre dann geschachtelt ('nested') in der Facette Prüfungssituation. Bei unvollständigen Designs ändern sich die oben angeführte Strukturgleichung sowie die Zusammensetzung der beobachteten Varianz entsprechend (Cornfield & Tukey, 1956). Nur bei vollständigen Designs lassen sich die einzelnen Fehlerkomponenten vollständig isolieren und getrennt schätzen, im unvollständigen Design lassen sich

die Interaktionen zwischen den geschachtelten Facetten nicht mehr isolieren (Cronbach et al., 1972, S. 45ff.).

Die Vorgehensweise bei der Ermittlung der Varianzkomponenten in einer G-Studie soll anhand des Prüfungsbeispiels demonstriert werden[1]. 100 Studierende (*p*) wurden im Laufe ihres Studiums zu zehn verschiedenen Prüfungszeitpunkten (verschiedene Stunden eines Tages, über eine Arbeitswoche verteilt) (*i*) mit drei verschiedenen Prüfungsformen (mündliche Prüfung, schriftliche Prüfung mit offenen und schriftliche Prüfung mit standardisierten Fragen) (*j*) geprüft. Man hat es mit zufälligen Facetten zu tun, über die hinaus generalisiert werden kann. Das Design ist vollständig, da jeder Studierende mit allen Prüfungsformen zu allen Zeitpunkten einmal geprüft wurde, wir haben also insgesamt 3000 Beobachtungen. Das Ergebnis der Varianzanalyse dieser hypothetischen G-Studie ist in Tabelle 1 wiedergegeben.

Tabelle 1: Ergebnisse der Varianzanalyse einer hypothetischen G-Studie zur Prüfung der Leistung von Studierenden

Facette	Quadratsumme	d.f.	Mittlere Quadratsumme
Studierender p	5999,40	99	60,60
Prüfungszeitpunkt i	1087,20	9	120,80
Prüfungsform j	6540,00	2	3270,00
p x i	1335,61	891	1,50
p x j	721,80	198	3,60
i x j	67,50	18	3,75
Fehler	2227,50	1782	1,25

Aufgrund der Zufallsauswahl der jeweiligen Bedingungen sind die beobachteten mittleren Quadratsummen in der Varianzanalyse unverzerrte Schätzer der Erwartungswerte der mittleren Quadratsumme ('expected mean square', kurz EMS). Diese EMS wiederum entstehen durch die Summe der gewichteten Varianzkomponenten, die wir ermitteln wollen (vgl. Cornfield & Tukey, 1956). Im oben genannten Beispiel ergeben sich für die Erwartungswerte die folgenden Aufsummierungen:

[1] Die Daten sind übernommen aus Peter (1979) und beziehen sich ursprünglich auf die Messung von Markentreue bei Konsumenten mit verschiedenen Items eines Fragebogens in verschiedenen Erhebungssituationen.

(6) $$EMS_p = \sigma_e^2 + n_i\sigma_{pj}^2 + n_j\sigma_{pi}^2 + n_i n_j \sigma_p^2$$

$$EMS_i = \sigma_e^2 + n_p\sigma_{ij}^2 + n_j\sigma_{pi}^2 + n_p n_j \sigma_i^2$$

$$EMS_j = \sigma_e^2 + n_i\sigma_{pj}^2 + n_p\sigma_{ij}^2 + n_p n_i \sigma_j^2$$

$$EMS_{pi} = \sigma_e^2 + n_j\sigma_{pi}^2$$

$$EMS_{pj} = \sigma_e^2 + n_i\sigma_{pj}^2$$

$$EMS_{ij} = \sigma_e^2 + n_p\sigma_{ij}^2$$

$$EMS_{res} = \sigma_e^2$$

Setzt man nun in der linken Seite die mittleren Quadratsummen ein, ist die unterste Gleichung bereits gelöst und durch Einsetzen dieser Lösung lassen sich die Gleichungen von unten nach oben lösen. Somit erhalten wird die geschätzten Varianzkomponenten für jede Facette, deren Interaktionen sowie den Fehler (vgl. Tabelle 2). Die Anteilsberechnung ist nun nicht mehr schwer: man addiert die einzelnen Varianzkomponenten auf und teilt jede einzelne Varianzkomponenten durch diese Summe. Für die prozentualen Anteile multipliziert man dann nochmals mit 100 (letzte Spalte in Tabelle 2).

Tabelle 2: Geschätzte Varianzkomponenten und Anteil an der Gesamtvarianz einer hypothetischen G-Studie zur Prüfung der Leistung von Studierenden

Facette	Mittlere Quadratsumme	d.f.	Schätzung der Varianzkomponente	Anteil an Gesamtvarianz
Studierender p	60,60	99	1,89	26,51
Prüfungszeitpunkt i	120,80	9	0,39	5,45
Prüfungsform j	3270,00	2	3,26	45,73
p x i	1,50	891	0,08	1,16
p x j	3,60	198	0,24	3,29
i x j	3,75	18	0,03	0,35
p x i x j, Fehler	1,25	1782	1,25	17,51
Summe			σ_i^2=7,138	100,00

Die Ergebnisse in Tabelle 2 verweisen auf einen hohen Beitrag zur Gesamtvarianz durch die Studierenden, die Prüfungsform und das Residuum, das mit der Dreifachinteraktion konfundiert ist. Als Facette der Differenzierung ist die Varianz bei den Studierenden erwünscht, allerdings zeigt der hohe Varianzanteil der Prüfungsform, dass

die geprüfte Leistung von Studierenden nicht ohne weiteres über verschiedene Prüfungsformen hinweg generalisiert werden kann. Weitere Facetten der Generalisierung, die nicht berücksichtigt wurden (z.B. Prüfungsfächer), können sich auch hinter der Varianzkomponente des Residuums verbergen.

An dieser Stelle sei noch anzumerken, dass es bei der Schätzung der Varianzkomponenten auf der Basis von Stichprobendaten insbesondere bei kleinen Stichproben zum Teil erhebliche Schätzfehler aufgrund von Stichprobenfehlern auftreten können (Shavelson & Webb, 1981). Dieser Problematik kann durch eine Vergrößerung der Stichprobe, eine Verringerung der Anzahl der Facetten oder eine Erhöhung der Zahl der Bedingungen je Facette entgegnet werden (Smith, 1978). Manchmal kommt es auch vor, dass aufgrund von Fehlspezifikationen oder von Messfehlern negative Werte für die Varianzkomponenten berechnet werden, die dann üblicherweise gleich Null gesetzt werden und dann entweder als Null oder aber als ursprünglicher negativer Wert in die weiteren Gleichungen zur Lösung der Erwartungswerte eingehen können (Brennan, 1983, S. 47f.; Shavelson & Webb, 1991, S. 37f.).

3.2.2 Optimierung eines generalisierbaren Studiendesigns in der D-Studie

In der G-Studie wird der Beitrag verschiedener Facetten zur Gesamtvarianz ermittelt. Diese Informationen werden in der Entscheidungs-Studie ('decision study', kurz D-Study) so verwendet, dass die Messung einen hohen Grad an Generalisierung erreicht. So wird ermittelt, welche Anzahl an Bedingungen für die einzelnen Facetten sinnvoll ist, um die Generalisierbarkeit zu erhöhen, d.h. den Anteil der Varianzkomponenten der Facette der Differenzierung möglichst groß und den Anteil der Facetten der Generalisierung möglichst klein werden zu lassen (Finn & Kayande, 1997; Nußbaum, 1987). Facetten der Generalisierung, die nur einen geringen Beitrag zur Gesamtvarianz liefern, können dabei auch ganz ausgeschlossen werden, da man offensichtlich von einer hinreichenden Generalisierung ausgehen kann. Das Studiendesign der D-Studie kann sich also von dem der G-Studie unterscheiden. Hat die G-Studie ein vollständiges Design, so kann die D-Studie auch ein unvollständiges Design verwenden oder mit festen Facetten arbeiten. Die G-Studie dient also insbesondere der Entwicklung eines generalisierbaren Messinstruments, die D-Studie optimiert dieses

Messinstrument dann zweckspezifisch. Dabei ist eine erneute Datenerhebung nicht nötig, vielmehr wird diese Optimierung durch die Berechnung von Varianzkomponenten und Generalisierbarkeitskoeffizienten für unterschiedliche Designalternativen erreicht.

Es können zwei Arten von Generalisierbarkeitskoeffizienten berechnet werden, abhängig davon, ob man dazu absolute oder relative Fehler verwendet (vgl. Abbildung 2). Relative Fehler umfassen alle Varianzkomponenten, die zum relativen Stand des interessierenden Untersuchungsobjekts bei einer Messung beitragen. Im Prüfungsbeispiel sind das die Interaktionen mit der Facette ‚Studierende' und das Residuum, nicht jedoch die Varianz der Prüfungsform und des Prüfungszeitpunkts, deren Effekte ja über alle Studierende gemittelt werden und daher keine Veränderung hinsichtlich der Relationen der Werte der einzelnen Studierenden auf der Leistungsskala mit sich bringen. Beim absoluten Fehler spielen diese Varianzkomponenten jedoch eine Rolle, denn je nach der Wahl des Prüfungszeitpunkts oder der Prüfungsform kann der absolute Wert einer Person auf der Skala der Prüfungsleistung anders ausfallen. Welcher Fehler nun zur Berechnung eines Generalisierbarkeitskoeffizienten zu verwenden ist, hängt also davon ab, ob man eine Entscheidung aufgrund absoluter Werte von Untersuchungsobjekten oder aufgrund einer Relation zwischen diesen Objekten zu treffen hat.

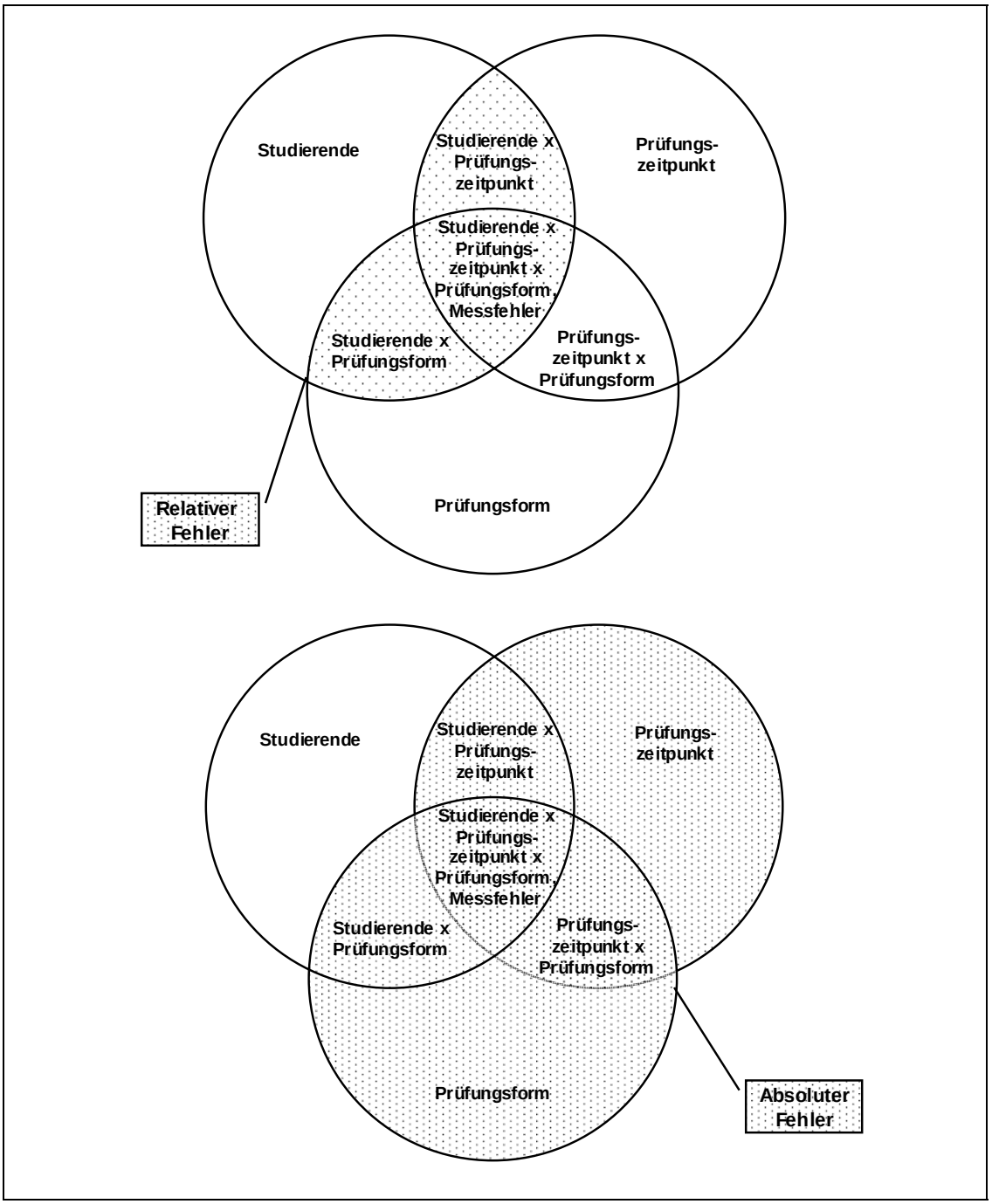

Abbildung 2: Relativer und absoluter Fehler der Messung

Analog zum Reliabilitätskoeffizienten der klassischen Testtheorie, der auf dem relativen Fehler beruht, berechnet sich ein Generalisierbarkeitskoeffizient $E\rho^2$ für relative Entscheidungen nun über die Varianz des 'universe score' der Person (oder allgemein der Facette der Differenzierung) im Verhältnis zur Summe aus der Varianz des 'universe score' und der relativen Fehlervarianz (Cronbach et al., 1972, S. 119ff.).

(7) $$E\rho^2 = \frac{\sigma_p^2}{(\sigma_p^2 + \sigma_{rel}^2)}$$

Für absolute Entscheidungen kann ein 'index of dependability' Φ dann entsprechend über das Verhältnis der Varianz des 'universe score' der Person (bzw. der Facette der Differenzierung) zur Summe aus der Varianz dieses 'universe score' und der absoluten Fehlervarianz berechnet werden (Brennan & Kane, 1977).

(8) $$\phi = \frac{\sigma_p^2}{(\sigma_p^2 + \sigma_{abs}^2)}$$

Ausgehend von zufälligen Facetten basiert der Generalisierbarkeitskoeffizient im Gegensatz zu den klassischen Reliabilitätskoeffizienten auf Erwartungswerten, wodurch wiederum eine Generalisierung des Koeffizienten über alle weiteren (nicht einbezogenen) Bedingungen einer Facette möglich ist (Rentz, 1987). Generalisierbarkeitskoeffizienten können in der Generalisierbarkeitstheorie relativ flexibel angewandt werden, insbesondere ist die Facette, für die der Generalisierbarkeitskoeffizient geschätzt werden soll, frei wählbar (Cardinet et al., 1976; Rentz, 1980).

Zur Optimierung des Studiendesigns anhand des Generalisierbarkeitskoeffizienten unterscheidet man meist zwischen zwei Zielkriterien. Entweder man maximiert bei einer festgelegten Anzahl von Beobachtungen (bzw. vorgegebenen Kosten einer Untersuchung) den Generalisierbarkeitskoeffizient durch die Variation der Bedingungen je Facette (z.B. durch die Veränderung der Anzahl der verwendeten Items). Oder aber man versucht mit möglichst wenig Beobachtungen (oder geringen Kosten) einen vorgegebenen Generalisierbarkeitskoeffizienten zu erreichen; in der Regel sind das Werte ab 0,9 (Finn & Kayande, 1997; Nußbaum, 1980, S. 106ff.). Es bleibt noch anzumerken, dass auch die Generalisierbarkeitskoeffizienten auf der Basis der geschätzten Varianzkomponenten ermittelt werden. Daher treten auch hier Stichprobenfehler auf, die typischerweise eher zu konservativen Schätzungen der Koeffizienten führen (Carroll & Faden, 1978).

3.3 Erweiterungen und Grenzen der Generalisierbarkeitstheorie

Soll neben der Ermittlung von Varianzkomponenten und Generalisierbarkeitskoeffizienten außerdem auch das Ergebnis der Messung in Form des 'universe score', also

des globalen wahren Wertes, ermittelt werden, so kann auf der Basis eines Beobachtungswerts, der ja annahmegemäß zufällig aus dem Universum zulässiger Beobachtungen gezogen wurde, ein Konfidenzintervall für den 'universe score' angegeben werden (Cronbach et al., 1972, S. 130ff.). Alternativ kann der 'universe score' einer Person (der Facette der Differenzierung) auch über eine Regressionsgleichung mit folgender allgemeiner Form ermittelt werden:

$$(9) \qquad \hat{\mu}_p = \mu + \rho^2(\overline{X}_p, \mu_p)(\overline{X}_p - E\overline{X}_p)$$

Die Konstante ist der Gesamtmittelwert μ, der über alle Facetten und deren Bedingungen aus den Daten, die der G- oder D-Studie zugrunde liegen, ermittelt wird. $\overline{X}_p$ ist der mittlere Beobachtungswert einer Person über alle weiteren Bedingungen, ermittelt auf der Datenbasis der D-Studie und der Regressionskoeffizient ρ^2 stellt die quadrierte Korrelation zwischen beobachteten Werten und 'universe scores' der Person für die gesamte Population (der Personen) und damit auch für das gesamte Universum zulässiger Beobachtungen dar. Dabei erweist sich insbesondere die Schätzung von ρ^2 bei einem für die Generalisierbarkeitstheorie kennzeichnenden Verzicht auf die Äquivalenzannahmen der Testtheorie als problematisch, weshalb man die ermittelten Schätzwerte auch eher als approximativ auffassen sollte (Brennan, 1983, S. 110; weiterführend vgl. Cronbach et al., 1972, S. 142ff.).

Die Generalisierbarkeitstheorie bietet auch die Möglichkeit der Analyse multivariater Fälle, also der Berücksichtigung mehrerer 'universe scores' eines Untersuchungsobjekts, so etwa wenn die "generelle" Markentreue gegenüber Getränkemarken ermittelt wird und dabei jeweils eine Fruchtsaftmarke, Biermarke und Limonadenmarke einbezogen wird. Dabei werden die Werte in Form eines Profils arrangiert oder zu einem zusammengesetzten Wert kombiniert, wobei neben den Varianzen der einzelnen Werte auch deren Kovarianzen bei der Auswertung berücksichtigt werden (weiterführend vgl. Brennan, 2001, S. 267ff.; Nußbaum, 1984; Shavelson & Webb, 1981).

Weiterhin sei noch darauf verwiesen, dass neben der Varianzanalyse auch andere Methoden zur Schätzung der Varianzkomponenten diskutiert werden, insbesondere Maximum-Likelihood-Schätzungen, die den Vorteil besitzen, dass negative Werte

für die Varianzkomponenten vermieden werden (vgl. auch die Übersicht über einzelne Schätzverfahren bei Brennan, 2001, S. 241ff.).

Schließlich soll noch die Nähe des Konzepts der Generalisierbarkeit zum Konzept der Validität angesprochen werden. Dabei werden unterschiedliche Bezugspunkte diskutiert (vgl. Cronbach et al., 1972, S. 378ff.; Kane, 1982). Am offensichtlichsten scheint dabei der Bezug zur Konstruktvalidität. Wird nämlich die Messmethode als Facette mit einbezogen, so wird über die Varianzkomponente dieser Facette gleichzeitig auch ein Maß für die Konvergenzvalidität eines Konstrukts geliefert; werden dagegen mehrere Teilaspekte eines Konstrukts untersucht oder aber im multivariaten Fall verschiedene Konstrukte , stellt der Wert einen Indikator für die Diskriminanzvalidität dieser Konstruktdimensionen bzw. Konstrukte dar (Campbell, 1976; Nußbaum, 1980, S. 105f.; Rentz, 1987).

Als die wichtigsten methodischen Probleme gelten die zum Teil erheblichen Stichprobenfehler bei der Schätzung der Varianzkomponenten sowie die Gefahr, bei dieser Schätzung negative Werte zu erhalten (Shavelson & Webb, 1981). Problematisch ist auch der Umgang mit unbalancierten Designs, also einer ungleichen Anzahl von Beobachtungen bei einzelnen Bedingungen, etwa wenn verschiedene Beobachter bei einem unvollständigen Design unterschiedlich viele Beobachtungen vornehmen, was ja eine grundsätzliche Problematik bei der Anwendung der Varianzanalyse darstellt (Searle, 1971, S. 35ff.). Peter (1979) weist weiterhin auf zwei praktische Beschränkungen des Ansatzes hin. Zum einen können das Design und auch die Interpretation insbesondere von Interaktionen höherer Ordnung bei der Berücksichtigung mehrerer Facetten sehr komplex werden. Wenn sie allerdings sinnvoll interpretierbar sind, stellen gerade auch die Interaktionen von Messfehlern einen entscheidenden Mehrwert der Generalisierbarkeitstheorie dar (Rentz, 1987). Zum anderen stellt sich die Frage der Effizienz einer meist doch recht aufwändigen Generalisierbarkeitsstudie. Um z.B. den Stichprobenfehler der Varianzkomponenten so klein zu halten, dass die Komponenten sinnvoll interpretiert werden können, empfiehlt Smith, (1978) bei zwei Facetten mindestens 800 Beobachtungen. Je umfangreicher die Definition des Universums, desto höher also der entsprechende Erhebungsaufwand (Cronbach et al., 1972, S. 372f.). Meist sind aber gerade die gewählten Items am stärksten für die Fehlervarianz verantwortlich, was aber genauso gut durch den Test der internen Konsis-

tenz im Rahmen der klassischen Testtheorie erfasst wird (Nunally, 1967, S. 210f.). Rentz, (1987) verweist darauf, dass häufig auch die Messbedingungen, die in der klassischen Testtheorie durch den Faktor Zeit berücksichtigt werden können, eine weitere wichtige Fehlerquelle darstellen.

4 Anwendungen der Generalisierbarkeitstheorie in der betriebswirtschaftlichen Forschung

In der nachfolgenden Tabelle findet sich eine Übersicht über Studien aus der betriebswirtschaftlichen Forschung, die Konstrukte anhand der Generalisierbarkeitstheorie untersuchen. Die Zusammenstellung basiert auf einer Datenbankrecherche in den Datenbanken Business Source Elite und ABI/Inform. Sie erhebt daher keinen Anspruch auf vollständige Abdeckung aller Generalisierbarkeitsstudien in der betriebswirtschaftlichen Forschung. Insbesondere ist die Übersicht auch auf publizierte Quellen aus englischsprachigen Zeitschriften beschränkt.

Folgende Punkte lassen sich festhalten:

- Anwendungen der Generalisierbarkeitstheorie in der betriebswirtschaftlichen Forschung finden sich vor allem in verhaltenswissenschaftlich und empirisch orientierten Forschungsgebieten wie der Personalforschung oder dem Marketing. Das liegt auf der Hand, da hier die Entwicklung von Messinstrumenten eine wichtige Rolle spielt. Allerdings zeigen Beispiele wie die Risikoeinschätzung von Versicherungsnehmern, dass die Generalisierbarkeitstheorie auch für darüber hinausgehende Forschungsgebiete und Themen von Interesse sein kann.
- Die meisten Anwendungen stellen Erweiterungen der klassischen Testtheorie dar und untersuchen Studienteilnehmer als Facette der Differenzierung und mehr als eine typische Fehlerquellen der Messung, wie wir sie aus der klassischen Testtheorie kennen (z.B. Items, Beurteiler) als Facette der Generalisierung.
- Einige Studien erweitern die klassische messtheoretische Herangehensweise, indem sie über Aspekte generalisieren, die in der klassischen Testtheorie nicht erfasst werden (z.B. Länder).

- Einige wenige Studien erfassen auch andere Facetten der Differenzierung als Personen, die die Generalisierbarkeitstheorie ja explizit zulässt (z.B. Werbungen, Geschäfte, Länder).
- Eher selten werden die Facetten der Differenzierung alterniert, so z.B. wenn einmal zwischen Werbungen oder Konsumenten als Facette der Differenzierung bei der Beurteilung von Werbungen unterschieden wird.
- Die meisten Studien ermitteln Generalisierbarkeitskoeffizienten und leiten daraus optimierte Studiendesigns ab.

Die Anwendungen zeigen neben der grundsätzlichen Brauchbarkeit der Generalisierbarkeitstheorie als messtheoretischen Ansatz auch deren Notwendigkeit auf, insbesondere dann, wenn es mehr als eine Fehlerquelle zu berücksichtigen gibt, beispielsweise wenn mehrere Beurteiler mit verschiedenen Items ein Urteil abgeben.

Dabei scheint das Potenzial der Generalisierbarkeitstheorie aber noch nicht völlig ausgeschöpft. Weitere Einsatzmöglichkeiten bieten sich an, wie z.B.:

- Es können andere mögliche Objekte der Differenzierung untersucht werden, z.B. Produkte, Firmen ebenso wie weitere Facetten der Generalisierung betrachtet werden können.
- Es kann weiter von der strengen Überprüfung des Messkonzepts abstrahiert werden, indem z.B. Fragestellungen wie die Problematik der internationalen Vergleichbarkeit von Ergebnissen untersucht werden, ohne dabei die Messung explizit zu fokussieren. So können z.B. Produktbeurteilungen von Konsumenten in verschiedenen Ländern untersucht werden und entschieden werden, ob sich dabei eher über Länder oder eher über Produkte generalisieren lässt.

Grundsätzlich wäre eine Berücksichtigung der Generalisierbarkeitstheorie als Standardverfahren bei Messtechniken, bei denen mehr als eine Fehlerquelle zu unterstellen ist, sinnvoll. Jedes Kodier- oder Begutachtungsverfahren, das mit mehr als einem Kodierer oder Gutachter und mehreren Items arbeitet, zählt dazu. Auf diese Weise ist beispielsweise auch die Generalisierbarkeit von Peer-Review-Verfahren bei der Begutachtung von wissenschaftlichen Arbeiten auf der Basis von standardisierten Beurteilungen messbar (Yavas, 1990).

Vor allem aber stellt die Generalisierbarkeitstheorie eine sinnvolle Ergänzung bisheriger Verfahren zu Sicherstellung der Qualität von Messungen dar, das in Kombination mit diesen anwendbar ist (Durvasula, Netemeyer, Andrews & Lysonski, 2006; Sharma & Weathers, 2003). Besser generalisierbare Merkmalsmessungen tragen dann auch zur Verbesserung der Generalisierbarkeit der Beziehungen von Merkmalen bei, indem nämlich der Generalisierungsbereich von der Merkmalsebene auch auf die Ebene der Merkmalsbeziehungen übertragen werden kann (Finn & Kayande, 1997; Morrison & Silva-Risso, 1995). Werden generalisierbare Merkmalsmessungen entwickelt, können diese auch bei Replikationsstudien eingesetzt werden und ermöglichen so eine bessere Vergleichbarkeit dieser Untersuchungen, da substantiell perfekte Messungen vorliegen (Hunter, 2001).

Tabelle 3: Übersicht über Anwendungen der Generalisierbarkeitstheorie in der betriebswirtschaftlichen Forschung

Quelle	*untersuchtes Konstrukt / Messkonzept*	*Facette der Generalisierung*	*Facette der Differenzierung*
Arthur, Woehr & Maldegen, 2000	Leistungsbewertung im Assessment Center	a) Teilnehmer b) Bewertungsdimensionen	- Aufgaben - Gutachter - a) Bewertungsdimensionen b) Teilnehmer
Burrows, Yavas, Bildici & Schweig, 1997 Burrows & Yavas, 1994	Risikoseinschätzung von Versicherungsnehmern	Versicherungsnehmer	- Versicherungsvertreter/Analysten - Beurteilungskriterien
DeSimone, Alexander & Cronshaw, 1986	Variabilität des finanziellen Werts der Leistung von Angestellten	Variabilität des finanziellen Werts auf drei Leistungsstufen	- Zeitpunkte - Beurteiler (Supervisor)
Deutskens, de Jong, de Ruyter & Wetzels, 2006	Servicequalität (SERVQUAL)	Länder	- Geschäftseinheiten - Personen - Qualitätsdimensionen - Items
Doverspike, Carlisi, Barrett & Alexander, 1983	Berufsbeurteilung	Berufe	- Beurteiler - Skalen (Beurteilungsdimension)
Durvasula et al., 2006	Generelle Einstellung zur Werbung	Teilnehmer	- Länder - Dimensionen - Items
Finn, 2001 - Studie 1	Qualitätsbeurteilung von Verkaufspersonal in Handelsketten	Handelsketten	- Mystery shopper - Dimensionen - Items

- Studie 2	Beurteilung der Ladenumgebung bei Handelsketten	Handelsketten	- Mystery shopper - Dimensionen - Items
- Studie 3	Qualitätsbeurteilung von Verkaufspersonal im einzelnen Laden	Einzelhandelsgeschäfte	- Mystery shopper - Dimensionen - Items
- Studie 4	Beurteilung der Ladenumgebung im einzelnen Laden	Einzelhandelsgeschäfte	- Mystery shopper - Dimensionen - Items
Finn, 2004	Servicequalität (SERVQUAL) bei Einzelhandelsgeschäften	Einzelhandelsgeschäfte und Mystery Shopper (multivariate G-Studie)	- Items - Dimensionen
Finn & Kayande, 1997	Servicequalität (SERVQUAL) bei Einzelhandelsgeschäften	Filialen (Handelssektoren)	- Qualitätsdimensionen - Items - Kunden
Finn & Kayande, 1999 - Studie 1	Servicequalität (SERVQUAL) bei Einzelhandelsgeschäften	a) Einzelhandelsgeschäfte b) Kunden	- Qualitätsdimensionen - Items - a) Kunden b) Einzelhandelsgeschäfte
- Studie 2	Servicequalität (SERVQUAL) bei Einzelhandelsgeschäften	a) Einzelhandelsgeschäfte b) Mystery Shopper	- Besuche - Qualitätsdimensionen - Items - a) Mystery Shopper b) Einzelhandelsgeschäfte
- Studie 3	Servicequalität (SERVQUAL) bei Einzelhandelsgeschäften, gemessen durch Mystery Shoppern	a) Einzelhandelsgeschäfte b) mystery shopper	- Items - a) Mystery Shopper b) Einzelhandelsgeschäfte
- Studie 4a	Servicequalität (SERVQUAL) bei Einzelhandelsgeschäften	a) Einzelhandelsgeschäfte b) Mystery Shopper	- Qualitätsdimensionen - Items - a) Mystery Shopper b) Einzelhandelsgeschäfte
- Studie 4b	Beurteilung der Ladenumgebung bei Einzelhandelsgeschäften	a) Einzelhandelsgeschäfte b) Mystery Shopper	- Qualitätsdimensionen - Items - a) Mystery Shopper b) Einzelhandelsgeschäfte
Gerhart, Wright & McMahan, 2000; Gerhart, Wright, McMahan & Snell, 2000	Bewertung der Personalpolitik	Unternehmen	- Beurteiler - Items
Greguras & Robie, 1998	Bewertung von Managern (360-Grad-Bewertung)	Manager	- Items - Gutachtertypen (Vorgesetzte, Kollegen, Mitarbeiter)

Greguras, Robie, Schleicher & Goff III, 2003	Leistungsbewertung im Unternehmen	Mitarbeiter	- Beurteiler - Item - Gutachtertyp (Vorgesetzter, Mitarbeiter)
Jackson, Stillman & Atkins, 2005	Leistungsbewertung von Teilnehmern im Assessment Center	Teilnehmer	- Aufgaben - Einzelaufgabe bzw. Personeneigenschaften
Komaki, Zlotnick & Jensen, 1986	Managerverhalten	Manager	- Beobachtungszeitpunkte - Aufgabendimensionen
Kraiger & Teachout, 1990	Arbeitsleistung	Angestellte	- Beurteiler - Items - Bezugsklassen (Dimension)
Muncy & Gomes, 1992	Beurteilung von Werbung	a) Einzelwerbungen b) Personen	- Items - a) Personen b) Einzelwerbungen
Robie, Born & Schmit, 2001	Persönlichkeitsdimensionen von Bewerbern	Bewerber	- Items - Situationen/Itemtypen
Webb, Shavelson, Shea & Morello, 1981	Einstellungstest	Berufsgruppen	- Beurteiler - Situationen - Lokalitäten
Sharma & Weathers, 2003	Ethnozentrismus von Konsumenten	Länder	- Konsumenten - Items
Yavas, 1989	Länderrisikoeinschätzung	Länder	- Analysten - Items

5 Zusammenfassende Bewertung und Ausblick

Die Generalisierbarkeitstheorie liefert einen wichtigen Beitrag für die Generalisierung von Merkmalsmessungen über vorher festgelegte Dimensionen der Generalisierung. Sie erlaubt eine flexiblere Anwendung als die klassische Testtheorie und scheint vor allem dann empfehlenswert, wenn mehr als eine bestimmte Fehlerquelle der Messung zu unterstellen ist. Damit ergänzt dieser Ansatz auf sinnvolle Weise bisherige Verfahren zur Sicherstellung der Güte von Messungen. Weiterhin lassen sich auch Fragen, die über die rein messtheoretischen Aspekte hinausgehen, beantworten, z.B. die Frage nach der Generalisierung von Ergebnissen über Länder hinweg. Durch die Möglichkeiten im Rahmen der D-Studien lassen sich auch wichtige praktische Implikationen ableiten, da ja genaue Aussagen darüber gemacht werden, wie eine Studie zu gestalten ist (z.B. wie viele Teilnehmer nötig sind) um den Mess-

fehler zu reduzieren und einen hohen Grad an Generalisierbarkeit zu erreichen. Die zunehmende Standardisierung von Messinstrumenten in vielen empirischen Disziplinen und die zugleich zunehmende Anwendung dieser Messinstrumente über verschiedene Kontexte hinweg (in verschiedenen Ländern, zu verschiedenen Zeitpunkten) zeigen, dass die Generalisierbarkeitstheorie einen wichtigen Beitrag für zukünftige Forschungsvorhaben liefern kann.

Literatur

Arthur, W., Jr., Woehr, D. J., & Maldegen, R. (2000). Convergent and Discriminant Validity of Assessment Center Dimensions: A Conceptual and Empirical Reexamination of the Assessment Center Construct-Related Validity Paradox. Journal of Management, 26, 813-835.

Brennan, R. L. (1983). Elements of Generalizability Theory. Iowa City, IO: American College Testing Program.

Brennan, R. L. (2001). Generalizability Theory. New York: Springer.

Brennan, R. L., & Kane, M. T. (1977). An Index of Dependability for Mastery Tests. Journal of Educational Measurement, 14, 277-289.

Burrows, T. M., & Yavas, B. F. (1994). Using G-Theory to Improve Lending Decisions. Journal of Retail Banking, 16, 27-31.

Burrows, T. M., Yavas, B. F., Bildici, H., & Schweig, B. B. (1997). The Use of Generalizabilty Theory For Lifre Insurance Underwriting Decisions. American Business Review, 15, 75-84.

Campbell, J. P. (1976). Psychometric Theory. In M. D. Dunnette (Ed.) Handbook of Industrial and Organizational Psychology (185-222). Chicago: Rand McNally College Publishing Company.

Cardinet, J., Tourneur, Y., & Allal, L. (1981). Extensions of Generalizability Theory And Its Applications in Educational Measurement. Journal of Educational Measurement, 18, 183-204.

Cardinet, J., Tourneur, Y., & Allal, L. (1976). The Symmetry of Generalizability Theory: Applications to Educational Measurement. Journal of Educational Measurement, 13, 119-135.

Carroll, R. U., & Faden, V. B. (1978). Some Sampling Characteristics of Three Estimators of the Intraclass Correlation. Educational and Psychological Measurement, 38, 855-863.

Cornfield, J., & Tukey, J. W. (1956). Average Values of Mean Squares in Factorials. Annals of Mathematical Statistics, 27, 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: John Wiley.

Cronbach, L. J., Rajaratnam, N., & Gleser, C. G. (1963). Theory of Generalizability: A Liberalization of Reliability Theory. British Journal of Statistical Psychology, 16, 137-163.

DeSimone, R. L., Alexander, R. A., & Cronshaw, S. F. (1986). Accuracy and Reliabilty of SDy Estimates in Utility Analysis. Journal of Occupational Psychology, 59, 93-102.

Deutskens, E., de Jong, A., de Ruyter, K., & Wetzels, M. (2006). Comparing the Generalizability of Online and Mail Surveys in Cross-National Service Quality Research. Marketing Letters, 17, 119-136.

Doverspike, D., Carlisi, A. M., Barrett, G. V., & Alexander, R. A. (1983). Generalizability Analysis of a Point-Method Job Evaluation Instrument. Journal of Applied Psychology, 68, 476-483.

Durvasula, S., Netemeyer, R. G., Andrews, J. C., & Lysonski, S. (2006). Examining the Cross-National Applicability of Multi-Item, Multi-Dimensional Measures Using Generalizability Theory. Journal of International Business Studies, 37, 469-483.

Finn, A. (2001). Mystery Shopper Benchmarking of Durable-Goods Chains and Stores. Journal of Service Research, 3, 310-320.

Finn, A. (2004). A Reassessment of the Dimensionality of Retail Performance: A Multivariate Generalizability Theory Perspective. Journal of Retailing and Consumer Services, 11, 235-245.

Finn, A., & Kayande, U. (1997). Reliability Assessment and Optimization of Marketing Measurement. Journal of Marketing Research, 34, 262-275.

Finn, A., & Kayande, U. (1999). Unmasking a Phantom: A Psychometric Assessment of Mystery Shopping. Journal of Retailing, 75, 195-217.

Gerhart, B., Wright, P. M., & McMahan, G. C. (2000). Measurement Error in Research on the Human Resources and Firm Performance Relationship: Further Evidence and Analysis. Personnel Psychology, 53, 855-872.

Gerhart, B., Wright, P. M., McMahan, G. C., & Snell, S. A. (2000). Measurement Error in Research on Human Resources and Firm Performance: How Much Error Is There and How Does It Influence Effect Size Estimates? Personnel Psychology, 53, 803-834.

Gleser, G. C., Cronbach, L. J., & Rajaratnam, N. (1965). Generalizability of Scores Influenced by Multiple Sources of Variance. Psychometrika, 30, 395-418.

Greguras, G. J., & Robie, C. (1998). A New Look at Within-Source Interrater Reliability of 360-Degree Feedback Ratings. Journal of Applied Psychology, 83, 960-968.

Greguras, G. J., Robie, C., Schleicher, D. J., & Goff III, M. (2003). A Field Study of the Effects of Rating Purpose on the Quality of Multisource Ratings. Personnel Psychology, 56, 1-21.

Hughes, M. A., & Garrett, D. E. (1988). Inter-Coder Reliability Estimation Approaches in Marketing: A Generalizability Theory Framework for Quantitative Data. Journal of Marketing Research, 27, 185-195.

Hunter, J. E. (2001). The Desperate Need for Replications. Journal of Consumer Research, 28, 149-158.

Jackson, D. J. R., Stillman, J. A., & Atkins, S. G. (2005). Rating Tasks Versus Dimensions in Assessment Centers: A Psychometric Comparison. Human Performance, 18, 213-241.

Kane, M. T. (1982). A Sampling Model for Validity. Applied Psychological Measurement, 6, 125-160.

Komaki, J. L., Zlotnick, S., & Jensen, M. (1986). Development of an Operant-Based Taxonomy and Observational Index of Supervisory Behavior. Journal of Applied Psychology, 71, 260-269.

Kraiger, K., & Teachout, M. S. (1990). Generalizability Theory as Construct-Related Evidence of the Validity of Job Performance Ratings. Human Performance, 3, 19-35.

Morrison, D. G., & Silva-Risso, J. (1995). A Latent Look At Empirical Generalizations. Marketing Science, 14, G61-G70.

Muncy, J. A., & Gomes, R. (1992). The Development of Advertising-Centered Versus Individual-Centered Scales. Journal of Current Issues and Research in Advertising, 14, 59-66.

Nunally, J. C. (1967). Psychometric Theory. New York: McGraw-Hill.

Nußbaum, A. (1987). Das Modell der Generalisierbarkeitstheorie. In K. J. Klauer (Ed.) Kriteriumsorientierte Tests. Lehrbuch der Theorie und Praxis lehrzielorientierten Messens (114-136). Göttingen: Hogrefe.

Nußbaum, A. (1980). Konstruktion, Planung und Analyse Lehrzielorientierter Tests auf der Grundlage der Generalisierbarkeitstheorie. Unveröffentlichte Dissertation. Technische Hochschule Aachen.

Nußbaum, A. (1984). Multivariate Generalizability Theory in Educational Measurement. An Empirical Study. Applied Psychological Measurement, 8, 219-230.

Peter, P. J. (1979). Reliability: A Review of Psychometric Basics and Recent Marketing Practices. Journal of Marketing Research, 16, 6-17.

Rajaratnam, N., Cronbach, L. J., & Gleser, G. C. (1965). Generalizability of Stratified-Parallel Tests. Psychometrika, 30, 39-56.

Rentz, J. O. (1987). Generalizability Theory: A Comprehensive Method for Assessing and Improving the Dependability of Marketing Measures. Journal of Marketing Research, 24, 19-28.

Rentz, R. R. (1980). Rules of Thumb for Estimating Reliability Coefficients Using Generalizability Theory. Educational and Psychological Measurement, 40, 575-592.

Robie, C., Born, M. P., & Schmit, M. J. (2001). Personal And Situational Determinants of Personality Responses: A Partial Reanalysis And Reinterpretation of the Schmit et al. (1995) Data. Journal of Business and Psychology, 16, 101-117.

Searle, S. R. (1971). Linear Models. New York: Wiley.

Sharma, S., & Weathers, D. (2003). Assessing Generalizability of Scales Used in Cross-National Research. International Journal of Research in Marketing, 20, 287-295.

Shavelson, R. J., & Webb, N. M. (1991). Generalizability Theory. A Primer. Newbury Park, CA: Sage.

Shavelson, R. J., & Webb, N. M. (1981). Generalizability Theory: 1973-1980. British Journal of Mathematical and Statistical Psychology, 34, 133-166.

Smith, P. L. (1978). Sampling Errors of Variance Components in Small Sample Multifacet Generalizability Studies. Journal of Educational Statistics, 3, 319-346.

Webb, N. M., Shavelson, R. J., Shea, J., & Morello, E. (1981). Generalizability of General Education Development Ratings of Jobs in the United States. Journal of Applied Psychology, 66, 186-192.

Yavas, B. F. (1989). An Exploratory Assessment of the Use of Generalizability Theory in Improving Country Risk Analyses. The Mid-Atlantic Journal of Business, 25, 51-61.

Yavas, B. F. (1990). The Generalizability of Measurements: An Application to "Blind Refereeing" Process. In B. J. Dunlap (Ed.) Developments in Marketing Science (289-293). Cullowhee, NC: Academy of Marketing Science.

Diskussionsbeiträge
des Fachbereichs Wirtschaftswissenschaft
der Freien Universität Berlin

2007

2007/1 BESTER, Helmut / Daniel KRÄHMER
Delegation and Incentives
Volkswirtschaftliche Reihe

2007/2 CORNEO, Giacomo / Olivier Jeanne
Symbolic Values, Occupational Choice, and Economic Development
Volkswirtschaftliche Reihe

2007/3 NITSCH, Volker
State Visits and International Trade
Volkswirtschaftliche Reihe