

## Zur Qualität von Methodentests: Eine kritische Stellungnahme zur Studie von Hoppe & Lamp (p&a 3/2001)



André Hahn ist Diplom-Psychologe und arbeitet als Senior Research Manager im Bereich Marketing Research der Pixelpark AG. Bevor er vor zwei Jahren die Verantwortung für die Online-Forschung und Toolentwicklung übernahm, arbeitete er neun Jahre als wissenschaftlicher Mitarbeiter an der Freien und der Humboldt-Universität in Berlin. Er ist Autor zahlreicher nationaler wie internationaler Publikationen.



Dr. Thomas Rodenhausen studierte Psychologie in Berlin. Nach seiner Promotion auf dem Gebiet der computergestützten Diagnostik arbeitete er zunächst als Wissenschaftlicher Assistent an der Freien Universität Berlin und wechselte dann zur MediaTransfer Netresearch und Consulting AG. Hier leitet er die Datenauswertung und Methodenentwicklung.

### Zusammenfassung

Die Evaluationsstudie von Hoppe & Lamp wird ihrem eigenen Anspruch, einen Beitrag

zur Validierung von Online-Marktforschung zu leisten, nicht gerecht. Die untersuchten Merkmale Rekrutierungs- und Befragungsart sind nicht nur miteinander untrennbar konfundiert, sondern mit einigen weiteren Merkmalen, so dass ein eindeutiger Rückschluss auf die Ursachen der festgestellten Unterschiede nicht möglich ist. Kritisiert wird weiter, dass die diskutierten Unterschiede im Brand Image weder im statistischen Sinne überzeugen, noch dazu geeignet sind, die drei Online-Erhebungsmethoden zu problematisieren. Insbesondere der nur am Rand erwähnte Befund der Unabhängigkeit der Interdependenzstruktur der Untersuchungsvariablen vom Untersuchungsansatz spricht für und nicht gegen die Qualität online erhobener Daten. Letztlich werden sowohl Analyseverfahren wie auch Indikatoren vorgeschlagen, die geeigneter wären, die Validität der Untersuchungsansätze zu prüfen.

Die Online-Marktforschung ist den Kinderschuhen entwachsen. Euphorische Debatten um die technischen und ökonomischen Möglichkeiten der Online-Forschung treten in den Hintergrund zugunsten des Ringens um Qualität im Sinne maximierter Studien-Validität. Abzulesen ist diese Entwicklung am starken Wachstum einschlägiger Publikationen (z.B. Theobald et al, 2001), dominierender Themen auf Tagungen wie der GOR 2001, der NetEffects oder dem 26. Kongress des BVM und den Inhalten von Arbeitskreisen wie NEON. In diesen Kontext ist auch die von Hoppe & Lamp in

Planung und Analyse 3/2001 vorgestellte Studie „Die Qualität von Online Panel – Ein Methodentest“ einzuordnen. Eine Intention der Autoren bestand darin, den Effekt von unterschiedlichen Rekrutierungsverfahren einerseits und von Erhebungsmethoden andererseits zu untersuchen. Da wir wesentliche Schlussfolgerungen und Interpretationen, die aus den Studienbefunden abgeleitet werden, nicht teilen, möchten wir in dieser Replik diskutierend auf die Untersuchung eingehen.

Die Studie von Hoppe & Lamp sucht die Frage zu beantworten, wie sich die Art der Rekrutierung sowie der Erhebung auf die Ergebnisse einer prototypischen Marktforschungsstudie auswirkt. Im Kern geht es um die Vergleichbarkeit der Studienergebnisse, wobei offensichtlich angenommen wird, dass die Feststellung mangelnder Vergleichbarkeit einen direkten Rückschluss auf die Validität der jeweiligen Untersuchungsform erlaubt. Natürlich ist dieser Schluss nur bedingt möglich, da die beiden untersuchten Merkmale allenfalls als notwendige, nicht aber als hinreichende Bedingungen der Validität von Studienergebnissen aus einer größeren Anzahl validitätsrelevanter Studienmerkmale (z.B. Reliabilität der Messung) identifiziert werden könnten. Die Studie könnte also auf zwei wichtige validitätseinschränkende Attribute aufmerksam machen, die jedoch umgekehrt die Validität von Untersuchungsdaten nicht sicher stellen.

Die Fragestellung von Hoppe & Lamp impliziert einen zweifaktoriellen Versuchsplan mit den

fünf Stufen des Gruppierungsfaktors Rekrutierungsart (vgl. Spalte 3 in Tabelle 1) und den zwei Stufen des Faktors Befragungsart (online versus offline). Hieraus ergeben sich insgesamt 10 Untersuchungsgruppen, von denen jedoch nur fünf tatsächlich realisiert wurden. Die beiden Faktoren sind also konfundiert. Selbst die beiden am ehesten zu vergleichenden Gruppen Online-Panel 1 und 2 unterscheiden sich nicht nur darin, ob sie online oder offline rekrutiert wurden, sondern noch zusätzlich in den Merkmalen Grundgesamtheit und Random-Sample (vgl. Tabelle 1). Online-Panel 1 wurde per Zufall, aktiv und offline aus der Grundgesamtheit der deutschen Wohnbevölkerung mit Telefonanschluss rekrutiert, wohingegen die Online-Rekrutierung von Online-Panel

2 nicht oder nur eingeschränkt zufällig sowie passiv aus einer nicht bekannten und schwer definierbaren Grundgesamtheit (z.B. Internet-Nutzer, die zwischen November und Dezember 2000 die Rekrutierungs-Websites besucht haben) erfolgte.

Tabelle 1 listet acht Merkmale, die über die fünf von Hoppe & Lamp untersuchten Gruppen mehr oder weniger systematisch variieren und somit als gleichberechtigte Faktoren zur Erklärung von Varianz in den abhängigen Variablen zwischen den Gruppen herangezogen werden können. Diese Merkmalsliste ist nicht vollständig und ließe sich fortsetzen - so variiert auch die Incentivierung der Teilnehmer über die Untersuchungsgruppen. Aus Sicht der erforderlichen kognitiven Informationsverar-

beitungsprozesse macht es möglicherweise auch einen Unterschied, ob die Fragen zur Beantwortung vorgelesen, auf einem Computermonitor dargestellt werden oder in Papierform zu bearbeiten sind. Unser Argument ist also ein einfaches kausalanalytisches: Der Untersuchungsansatz von Hoppe & Lamp ist grundsätzlich wenig geeignet, eindeutige Bedingungen oder gar Ursachen von festzustellenden Unterschieden oder auch fehlenden Unterschieden derart merkmalsheterogener Untersuchungsgruppen auszumachen. Jede Interpretation und Schlussfolgerung von Hoppe & Lamp ist daher nur eine von mehreren Möglichkeiten und hat nur Bestand unter Annahme wechselnder Rahmenbedingungen.

Tabelle 1

Gruppe	Rekrutierungsgrundgesamtheit	Kontaktaufnahme	Interviewereffekt	Interviewzeitpunkt	Befragungshäufigkeit	Rekrutierung	Auswahlverfahren	Befragungstyp
IPSOS CAPI-Bus	BRD 14-69	face-to-face	hoch	sofort	Erstbefragte	offline	random	offline
Online-Panel 1	BRD 14-69 mit Telefonanschluss	Telefon	gering	nach Wunsch	Erstbefragte	offline	random	online
Online-Panel 2	unbekannt bzw. nicht definiert	Online	sehr gering	nach Wunsch	Erstbefragte	online	non-random	online
IPSOS-Online-Panel	ACCESS-Panel (72.000)	Telefon	gering	nach Wunsch	Mehrfach-Befragte (offline 4-5 mal pro Jahr)	offline	non-random	online
IPSOS Access Panel	ACCESS-Panel (72.000)	postalisch	sehr gering	nach Wunsch	Mehrfach-Befragte (offline 4-5 mal pro Jahr)	offline	non-random	offline

Allein die Unterschiede bleiben und laden zu Recht zu spannenden Spekulationen ein. Der Warnung von Hoppe & Lamp vor Online-Befragungen (oder zumindest vor panelbasierten Online-Befragungen), können wir jedoch auf der Grundlage der berichteten Daten nicht folgen, selbst wenn wir das geschilderte multiple Konfundierungsproblem einfach ignorieren und uns auf die Unterschiede an sich konzentrieren - ohne ihre Ursachen klären zu wollen.

Relevante Unterschiede zwischen den verschiedenen Untersuchungsansätzen versuchen die Autoren anhand diverser Image-Indikatoren der Marke Magnum auszumachen, da insbesondere deren Validität vom Auftraggeber eingefordert wird. Geprüft werden zu diesem Zweck Unterschiede zwischen den fünf Gruppen im Imageprofil sowie Unterschiede in der Brand Health Map. Dem Text (S. 48, Spalte 3) und der Abbildung 1 ist aber überraschenderweise zu entnehmen, dass statistisch signi-

fikante univariate Unterschiede nur in drei Fällen aufgetreten sind: zwischen den postalisch befragten Access-Panelisten und allen anderen Gruppen. Unabhängig von der offenen Frage, ob diese drei univariaten Unterschiede auch in der multivariaten Analyse Bestand beweisen konnten, bedeutet dies im Umkehrschluss, dass sich die verbleibenden vier Gruppen bezüglich aller Image-Merkmale nicht paarweise voneinander unterscheiden. Von 100 non-redundanten möglichen paar-

weisen Vergleichen werden nur drei als überzufällig unterschiedlich ausgewiesen. Deshalb erschien es uns zunächst geboten, die positive Antworttendenz des IPSOS-Mailpanels zu diskutieren (Als Hypothesen böten sich hier z. B. die Übertestung der Stichprobe oder die stimmungsherbende Bearbeitung des attraktiven Langnese-Fragebogens im Familienverbund an.). Möglicherweise sind aber selbst die abweichenden Imagewerte des IPSOS-Mailpanels auf die Unterschiede der nicht-quotierten demographischen Merkmale oder der Merkmale des Internetnutzungsverhaltens zurückzuführen. Eine kovarianzanalytische Prüfung könnte hier Aufschluss geben. Eine solche Prüfung könnte die Frage beantworten, ob und in welcher Form die demographischen Unterschiede und die Unterschiede im Internetnutzungsverhalten tatsächlich die Imageeinschätzung mit beeinflussen - eine Hypothese, die auch von Hoppe & Lamp (S. 50, Spalte 1) aufgeworfen wird.

Die in der Brand Health Map (Abbildung 2) ausgewiesenen deutlicheren Unterschiede zwischen den Untersuchungsgruppen erscheinen in der graphischen Darstellung überzeugend. Allerdings ist die Brand Health Map redundant zum Brand Image Profil, da die Darstellung auf berechneten Indikatoren einer Auswahl der gleichen Variablen beruht. Zu sehen sind erhebliche Unterschiede, die vormals nicht zu erkennen waren. Nicht nachzuvollziehen ist, dass die Kreise der Einschätzung des Preis-Leistungsverhältnisses der beiden Offline-Studien doppelt so groß sind wie die der drei Online-Studien. Diese überdeutlichen Unterschiede widersprechen zumindest der Darstellung des Brand Image Profils in Abbildung 1. Im Vergleich zu den beim Brand-

Image Profil geschilderten Unterschieden tritt bei der Brand Health Map ein weiterer überraschender Unterschied auf: Die IPSOS CAPI-Bus-Studie hat gegenüber den drei Online-Studien eine höhere Ausprägung auf dem multiplikativ gebildeten Index „Power“ („diese Marke passt zu mir“ x „ist anders als andere Marken“). Hierfür lassen sich zwei Ursachen anführen: Zum einen wird durch die Multiplikation der Variablen eine größere Varianz des Indikators zugunsten der CAPI-Bus-Gruppe erzeugt, da beide Imagevariablen bereits vergleichsweise hoch bei dieser Gruppe ausgeprägt sind (vgl. Abbildung 1). Zum anderen könnte hier auch ein differentielles Messfehlerproblem vorliegen, da die Reliabilität des multiplikativen Indikators Power niedriger ist als die der Einzelindikatoren. Dies wäre genau dann der Fall, wenn sich die Varianzen der den Indikator Power bildenden Variablen der CAPI-Bus-Gruppe von den drei Online-Gruppen unterscheiden. In diesem Fall wäre einfach der Messfehler in der CAPI-Bus-Gruppe systematisch höher als in den drei Online-Gruppen. Diese Hypothese ist trotz unbekannter Reliabilitäten testbar. Fazit: Die Brand Health Map bestätigt im wesentlichen die Ergebnisse der (fehlenden) Unterschiede im Brand Image Profil.

Ein letztes wichtiges Ergebnis der Studie von Hoppe & Lamp ist zu diskutieren. Die Autoren stellen fest, dass die relative Position der Marken zueinander in der Brand Health Map von den fünf unterschiedlichen Untersuchungsansätzen nahezu unberührt bleibt (S. 49). Das ist aufgrund der Abbildung 4 sehr gut nachzuvollziehen, auch wenn uns nicht klar ist, warum Magnum hier abweichend und widersprüchlich zur Positionierung im Koordina-

tensystem in Abbildung 2 verortet ist. Dieser Befund spricht für eine unabhängig vom Untersuchungsansatz ähnliche Kovarianzstruktur zwischen den Untersuchungsvariablen. Diese Hypothese wäre durch einen Gruppenvergleich in einem Strukturgleichungsmodell prüfbar. Die Konsequenzen identischer Kovarianzstrukturen wären für die Online-Marktforschung erfreulich, denn dann ließen sich Interdependenzen verzerrungsfrei analysieren. So wäre beispielsweise die Analyse von Veränderungen, von experimentellen Versuchsplänen oder von Bedingungs-Folge-Zusammenhängen nicht von möglichen Methodeneffekten der Online-Forschung betroffen.

Die Evaluationsstudie von Hoppe & Lamp wird unseres Erachtens ihrem eigenen Anspruch, einen Beitrag zur Validierung von Online-Marktforschung zu leisten, nicht gerecht. Wenn die Qualität unterschiedlicher Rekrutierungstypen und Befragungsarten geprüft werden soll, warum werden keine Kriterien definiert, anhand derer die Datenqualität abzulesen ist? Gefragt sind hier Reliabilitäts- oder Validitätsparameter wie interne Konsistenzen oder Korrelationen mit externen Merkmalen. Der Frage uni- oder multivariater Stichprobenhomogenität wird analytisch nicht nachgegangen. Sind die Varianzen der Image-Merkmale über die Untersuchungsgruppen hinweg vergleichbar oder verschieden? Durchaus plausibel wäre beispielsweise die Annahme, dass die Online-Samples generell homogenere Gruppen bilden als die beiden Offline-Gruppen, was an Streuungsunterschieden ablesbar wäre. Oder: Wie sieht es mit der Ausprägung klassischer Antworttendenzen wie sozial erwünschtem Antwortverhalten aus? Werden online mehr "Chaos-Daten"

mehr "Chaos-Daten" produziert, treten häufiger zufällige Antwortmuster auf?  
Trotz aller Kritik möchten wir den Autoren darin beipflichten, Mindestanforderungen an die Qualität von Marktforschungsstudien - ob nun online oder offline – zu fordern und teilen auch, die eigentlich selbstverständliche Auffassung, dass nicht jede Fragestellung mit Hilfe von Online-Studien zu beantworten ist. Letztlich sollte die Wahl der Untersuchungsmethodik als ein Problem der Validitätsoptimierung unter Berücksichtigung ökonomischer Rahmenbedingungen aufgefasst werden.

#### Literatur

Theobald, A., Dreyer, M. & Starsetzki, T. (2001) (Hg.). Online-Marktforschung. Beiträge aus Wissenschaft und Praxis. Wiesbaden: Gabler.

Erscheint in: Planung & Analyse 4 / 2001 URL:  
<http://www.planung-analyse.de>

#### Kontakt

Dipl.-Psych. André Hahn  
Rotherstr. 8  
10245 Berlin  
Tel: + 49 30 5058 - 1841  
Fax: + 49 30 5058 – 1344  
Email: andre.hahn@pixelpark.com

Dr. Thomas Rodenhausen  
MediaTransfer AG  
Netresearch & Consulting  
Rothenbaumchaussee 38, 20148  
Hamburg  
Tel: (040) 669 625 0  
Fax: (040) 669 625 29  
Email:  
t.rodenhausen@mediatransfer.de