Kann Werbewirkung sensibler gemessen werden?

Dirk Frank und Wulf Schlund

„Werbepretests – Praxis und Erfolgsfaktoren“ heißt das neue Buch von Volker Trommsdorf. Der mit Abstand wichtigste „Erfolgsfaktor“ ist, dass der Pretest **vor** der Schaltung der Werbung durchgeführt wird, der zweitwichtigste die „Realitätsnähe des Werbemittels“. Praktiker würden hier eher von „Grundvoraussetzungen“ als von „Erfolgsfaktoren“ sprechen, denn zu spät durchgeführte Pretests sind in der Tat weit weg vom Ideal.

Trommsdorf untersucht das Pretestmanagement und die Faktoren, unter denen getestete Werbekampagnen erfolgreich sind. Auch wenn sich in der Praxis nicht alle seine Empfehlungen realisieren lassen – z. B. Werbeagenturen als Pretestdienstleister – gibt er doch einen bemerkenswerten Überblick über die gängige Pretestpraxis in Deutschland. Pretest**methoden** vergleicht das Buch in weiser Absicht nicht, sie waren auch nicht Gegenstand der Untersuchung.

1. Werbepretests bisher: wo kann überhaupt etwas verbessert werden?

Eines kann man ganz gewiss nicht sagen: nämlich, dass es zu wenige Pretestverfahren gäbe. Die existierenden Verfahren sind ausgeklügelt und untersuchen selbst die kleinsten Nuancen von Wirkungskriterien, teilweise bis zu absurden Details. So ließe sich aus manchen Instrumenten, die das Gefallen der musikalischen Elemente eines Werbespots untersuchen, eine regelrechte „Schlager der Woche“ -Sendung gestalten. Auch die Vielzahl der untersuchten Kriterien von Werbewirkung ist, nimmt man alle Verfahren zusammen, eher beängstigend groß als lückenhaft. Gibt es da überhaupt noch Felder, auf denen sich Schwächen oder Mängel und damit Verbesserungspotentiale ausmachen lassen? Bei genauem Hinsehen konnten wir drei Felder ausmachen, auf denen das der Fall ist:

1. Schaut man sich jedes einzelne Verfahren an, stellt man Lücken in der systematischen Verknüpfung von Kriterien fest. Insbesondere entdeckt man Defizite bei der psychologischen Validität bzw. Kongruenz der Kriterienerhebung. In der Pre-Post-Messung ist z. B. die Pre-Definition von Markenpräferenz über Kaufverhaltensabfragen oder Relevant Set mit der Post-Definition eines „Gewinnkorbs“ psychologisch nicht vergleichbar: Kaufverhalten hängt in der Realität auch von Preisstellung und Verfügbarkeit ab, die Bestückung von Gewinnkörben folgt dagegen anderen psychologischen Spielregeln (z. B. „Champagnereffekt“).

2. Ein weiteres Manko bestehender Pretestverfahren – und darauf hat Volker Trommsdorf, vor allem aber Ulrich Lachmann in seinem Buch „Wahrnehmung und Gestaltung von Werbung“ hingewiesen – ist die ungenügende Berücksichtigung der Rezeptionsbedingungen von Werbung, d. h. des unterschiedlichen Involvements bei der Wahrnehmung der Werbung.

3. Das dritte Manko von Pretests ist das bekannteste und am häufigsten kritisierte: ihre mangelnde Sensibilität. Dass Pretests nicht sensibel genug auf die kreative Leistung eines Kampagnenansatzes eingehen, wird grundsätzlich von den Werbeagenturen beklagt; dass sie mögliche Wirkungen des Werbemittels nicht oder nicht genügend deutlich anzeigen, auch vom Marketing der Werbungtreibenden. Mit diesen Argumenten wird auch nach wie vor oft die Skepsis gegenüber Pretests begründet, und sowohl positive Pretestergebnisse bei nachfolgendem Misserfolg im Markt als auch schlechte Pretestergebnisse mit nachfolgendem Markterfolg werden als Beispiele aufgeführt.

Wie weit diese Kritik wirklich berechtigt ist und in welchem Maße dieses immer wieder zitierte Manko tatsächlich Grund des Auseinanderklaffens von Pretestergebnissen und Marktwirklichkeit ist, lässt sich wohl kaum objektiv feststellen, da zwischen Pretest und Marktwirklichkeit immer Faktoren wirksam werden, die nicht kontrolliert werden. Aber das ist auch nicht nötig. Entscheidend bleibt die Frage, ob Werbewirkung sich präziser und sensibler erfassen lässt. Kann die Persuasionwirkung (z. B.

Shifts in der Markenpräferenz) sensibler gemessen werden? Sind die bisherigen Operationalisierungen immer noch zu grob? Insbesondere bei starken Marken, wo potentielle Werbeeinflüsse nicht sofort verhaltenswirksam werden? Und lässt sich die Messung selbst noch sensibler vornehmen?

Es sind diese Kernfragen, auf die dieser Beitrag eine Antwort geben möchte.

2. Können Pretests sensibler werden?

Operationalisierung und Messung:

Obwohl diese Frage im Prinzip für alle Kriterien der Werbewirkung gestellt werden kann, wollen wir uns in diesem Beitrag auf das besonders wichtige und heikle Kriterium der Persuasion-Wirkung konzentrieren: ausgelöstes Produktinteresse, Kaufbereitschaft, Shifts in der Markenaffinität. Bei der Untersuchung der Persuasion-Wirkung kann mangelnde Sensibilität an zu grober Operationalisierung oder an zu grober Messung liegen. Abgesehen von fehlender Kongruenz in der Pre-Post-Messung – hier Markenpräferenz über Kaufverhaltensabfragen oder Relevant Set-Ermittlung, dort Post-Definition eines „Gewinnkorbs“ – auch wenn der Gewinnkorb vorher und nachher eingesetzt würde, könnte es immer noch am Gewinnkorb liegen, wenn keine Bewegung zwischen Pre- und Post-Messung erkennbar wird. Eine andere Methode ist besser: Brand Choice Behavior Shifts (BCBS) werden am besten durch Chip Games gemessen, und zwar in der Pre- wie in der Post-Phase. Chip Games sind Gewinnkörben aus mehreren Gründen überlegen, insbesondere wenn man einfache Chip Games mit gestuften Paarvergleichsverfahren zwischen den Marken des Relevant Set kombiniert.

Kaufverhalten und Markenaffinität:

Persuasion muss aber nicht gleich bis auf die Verhaltensebene wirken. Auf vielen Produktfeldern, insbesondere auf solchen mit mehreren starken Marken, kann die Persuasion-Wirkung von Werbung zunächst nur subtiler einsetzen und nicht sofort aus Nicht-Käufern einer Marke Käufer machen. Sehr wohl kann Werbung in solchen Fällen aber die Markenaffinität eines Verbrauchers positiv beeinflussen, oft vielleicht nur latent. Es wird also aus einem gleichgültigen Verbraucher nicht sofort ein kaufinteressierter, sondern aus einem Ablehner ein nachdenklicher, aus einem nachdenklichen ein aufmerksamer, aus einem aufmerksamen ein interessierter Verbraucher, der aber trotzdem erst beim übernächsten Mal kauft. Welcher Pretest misst eigentlich diese Markenaffinität der Verbraucher, die kontinuierlich und nicht dichotom abgestuft ist und sich daher weder mit Kauf oder Nicht-Kauf noch mit anderen kategorialen Operationalisierungen einfangen lässt? Das Argument, solche subtilen Veränderungen der Markenaffinität seien keine hinreichende Werbewirkung, wäre naiv, denn nur selten führt ein einmaliger Werbeanstoß direkt zum Kauf.

Die Messung eines Kontinuums:

Dass Pretests subtile Shifts in der Markenaffinität, noch dazu im latenten Bereich, bisher nicht gemessen haben, ist kein Zufall, denn es gab dafür bisher auch kein wirklich geeignetes Instrument. Gerade wenn man sich die Markenaffinität eines Verbrauchers als stufenloses Kontinuum vorstellt, lässt sich dieses Kontinuum ja auf keiner der herkömmlichen Skalen adäquat repräsentieren. Aus diesem Grunde kommt der von ISM GLOBAL DYNAMICS neu entwickelten Stimulus Comparison Scale (SCS) besondere Bedeutung zu. Denn dieses neu entwickelte Messinstrument enthält im Unterschied zu herkömmlichen Skalen kein Grading, sondern stellt ein stufenloses Kontinuum dar, das deswegen auch die Markenaffinität der Verbraucher optimal operationalisiert. Auf diesem stufenlosen Kontinuum lässt sich auch für solche Marken, die nicht zum Kauf ausgewählt wurden, das latente Potential bei jeder Person ermitteln. Das bedeutet, dass mit der SCS-Skala auch solche Werbewirkungen identifiziert werden können, die sich gröberen Messungen bisher entzogen haben.

Das neue Messinstrument:

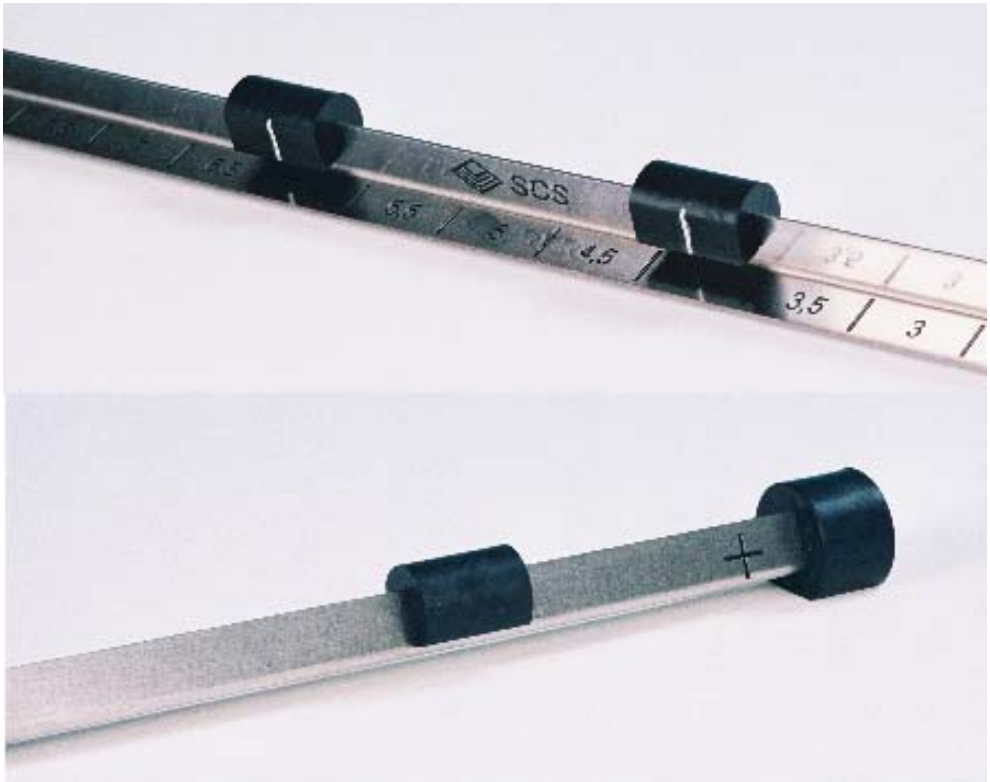
Die SCS stellt ein total neues Messinstrument dar, weil nämlich die stufenlose Skala lediglich eine Standardstrecke für stufenlose Einstufungen bietet. Die SCS erfüllt die psychometrische Maximalforderung der skalenfreien Anordnungen von Objekten mit der minimalen Einschränkung, dass die Gesamtstrecke, innerhalb derer Objekte miteinander verglichen werden, festgelegt ist. Diese Gesamtstrecke sowie ein Minus-Symbol auf der einen und ein Plus-Symbol auf der anderen Seite als Richtungsanzeiger der Einstufung sind die einzigen Vorgaben, die das neue Messinstrument der Testperson macht. Auf der SCS können Befragte ihre Bewertungen darstellen, ohne sie zuvor in verbale oder numerische Werte übersetzen zu müssen. Sie können damit freier und kontinuierlicher als auf jeder herkömmlichen Skala urteilen. Die messtheoretische Überlegenheit der SCS ist gravierend und für die ganze Marktforschung interessant:

1. Sie minimiert die Fehlereffekte, die bisher durch verbale, numerische oder räumliche Skaleneinteilungen entstanden sind: Halo-Effekte, Klumpungen, Tendenzen zu nivellierenden oder extremen Urteilen.

2. Durch das Wegfallen verbaler, numerischer und räumlicher Orientierungsanker können diese auch nicht mehr kulturell bedingte Unterschiede zwischen Verständnis und Handhabung zwischen verschiedenen Ländern auslösen. Das macht internationale Ergebnisse überhaupt erst wirklich vergleichbar.

Die messtheoretische Überlegenheit der SCS wurde inzwischen in mehreren Validierungsstudien bestätigt, für die Datenqualität an sich und auch für die Aufklärungskraft von Varianz, wie wir am Beispiel der Modellierung von Key Triggers zeigen.

Schaubild 1: Stimulus Comparison Scale (SCS)



a) Datenqualität

In 30 Produkt- und Konzepttests wurde die SCS mit einer 5-stufigen Verbalskala verglichen. Mit Hilfe einer entsprechenden Transformationsformel wurden dann die Antwortprofile beider Skalen gegenübergestellt. Schaubild 2 zeigt die bessere Qualität der SCS Messung deutlich. Das SCS-Antwortprofil zeigt eine klare Normalverteilungskurve, wie sie entsteht, wenn Produkte verschiedener Stilrichtungen von Testpersonen mit unterschiedlichem Stilempfinden beurteilt werden. Das Antwortprofil der 5er-Stufenskala weist dagegen deutliche Verzerrungen auf: während die positiven Urteile in vergleichbarer Häufigkeit auftreten, zeigen sich im mittleren und negativen Skalenbereich deutliche Abweichungen von der Normalverteilung.

b) Erklärte Varianz bei der Modellierung von Key Triggers

Gute Werbepretests erfassen nicht nur Veränderungen in der Produktwahrnehmung, sondern erklären sie auch. Sie stellen nicht nur sicher, dass die Wahr-

Schaubild 2: Antwortprofile von SCS und 5er-Stufenskala im Vergleich

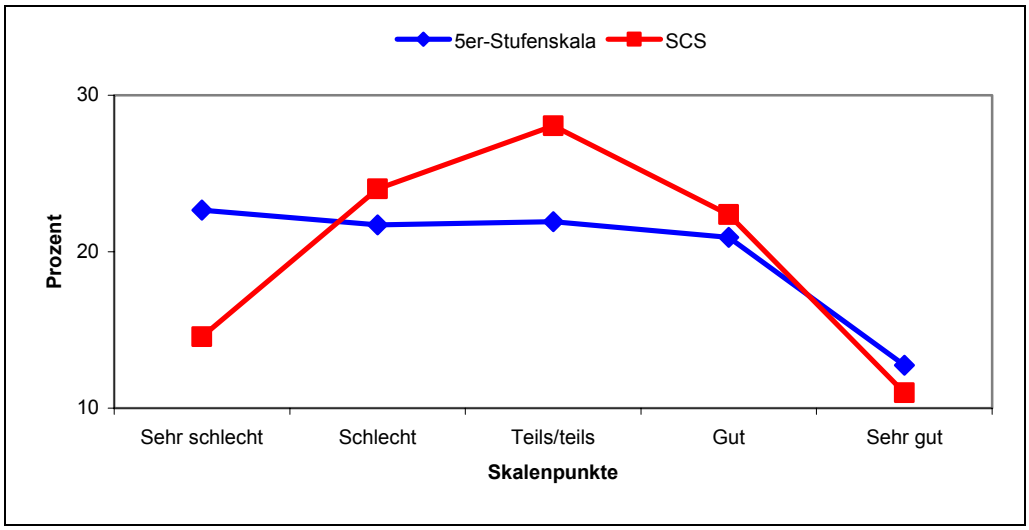
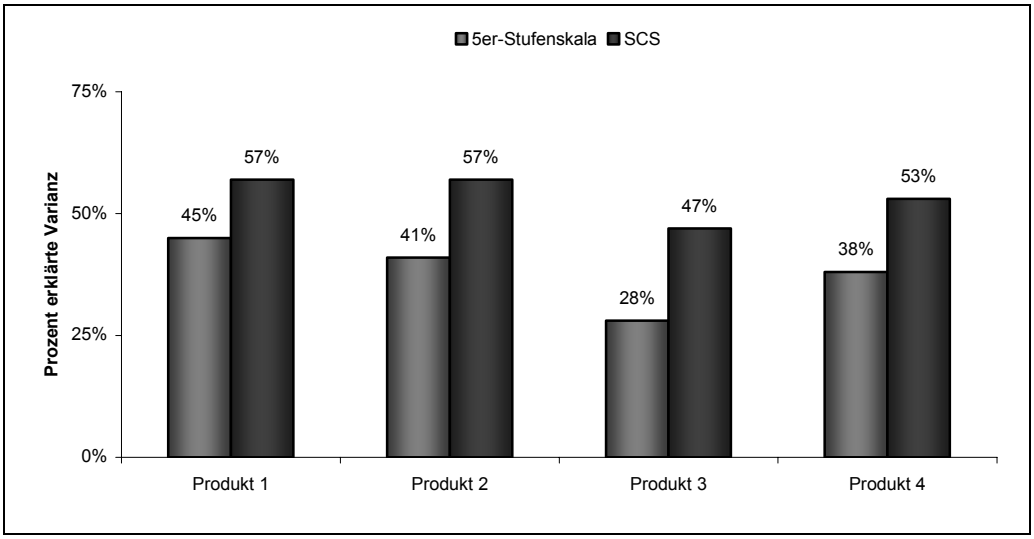


Schaubild 3: Analyse kaufrelevanter Fktoren mit SCS und 5er-Stufenskala im Vergleich



nehmungsänderung durch die Werbung ausgelöst wurde, sondern identifizieren auch diejenigen Kriterien, die den größten Einfluss auf die Veränderungen haben. Das geschieht mit der so genannten Key Trigger-Analyse, in der die Veränderung in der Markenaffinität als abhängige Variable durch Veränderungen in der Produkterwartung erklärt wird. Auch eine solche Analyse ist umso erfolgreicher, je höher die Qualität der Daten ist. Aus vier Pretests für Fertigprodukte wurden Key-Trigger-Analysen zur Erklärung von Kaufbereitschaftergebnissen durch Produkterwartungen durchgeführt: jeweils mit einer herkömmlichen 5er-Skala und der SCS. Das Ergebnis ist in Schaubild 4 dargestellt. Es zeigt, in welchem Ausmaße die skalierte Kaufbereitschaft durch die Summe der Produkterwartungen erklärt wird. In allen vier Fällen ist die Erklärungskraft bei der SCS deutlich höher als bei der 5er-Stufenskala, obwohl die Regressionsmodelle inhaltlich und technisch identisch waren. So wird durch die Verwendung eines präzisen Messinstrumentes auch die Analyse der kaufrelevanten Faktoren wesentlich erfolgreicher und verlässlicher.

3. Internationale Vergleichbarkeit

Im Globalisierungszeitalter ist die internationale Vergleichbarkeit von Ergebnissen noch wichtiger geworden. Und hier liegt die große Schwäche herkömmlicher Stufenskalen. Selbst wo die adäquate Übersetzung verbalisierter Skalenpunkte unproblematisch ist, weiß man, dass das bei Testpersonen verschiedener Länder unterschiedliche Assoziationen hervorruft. Wissen wir wirklich, ob eine italienische Testperson mit sehr gut das gleiche meint wie eine schwedische oder eine deutsche? Die Erfahrung zeigt eher das Gegenteil, und das schränkt die Vergleichbarkeit von Ergebnissen auf Stufenskalen ein, erst recht, wenn sie verbalisiert sind.

Genau hier liegt der ganz große Vorteil der SCS. Da die SCS auf Stufen und Verbalisierungen völlig verzichtet, können diese auch nicht mehr zu kulturell bedingten Unterschieden im Antwortverhalten führen. Eine Validierungsstudie sollte herausfinden, wie sich das in der Praxis tatsächlich bemerkbar macht.

Dazu haben wir einen Konzepttest für ein europaweit führendes Markenprodukt aus dem FMCG-Bereich benutzt, der vergleichend mit der SCS und einer herkömmlichen 5er-Stufenskala durchgeführt wurde. Die Schaubilder 4 und 5 zeigen die gewonnenen Antwortprofile für beide Skalen in Italien, Deutschland und Schweden.

Die Antwortprofile errechnen sich aus den gemittelten Häufigkeiten pro Skalenpunkt über alle skalierten Fragen, die sich auf das untersuchte Produkt beziehen. Die Skalenwerte der SCS wurden hierfür in die der 5er-Skala transformiert. Da beide Antwortprofile auf identischen Fragen zum selben Objekt beruhen, sollte man identische Kurvenverläufe erwarten, wenn beide Skalen die Produktbeurteilungen unverzerrt erheben. Treten jedoch Abweichungen zwischen Beurteilungen mit der SCS und der Stufenskala auf, sind sie ein Hinweis darauf, dass die Skalen unterschiedliches Antwortverhalten bewirken.

Schaubild 4 zeigt die Antwortprofile für die SCS, Schaubild 5 für die Stufenskala. Da sich alle Fragen in dieser Validierungsstudie auf ein Premiumprodukt beziehen, liegen fast alle Bewertungen im positiven Bereich. Die Kurven steigen daher nach rechts an und bilden keine Normalverteilung. Dieser Anstieg ist allerdings nur bei der SCS glatt und konsequent und zwischen den drei Ländern deutlich ähnlicher als bei der Stufenskala. Die Antwortprofile der Stufenskala zeigen ein ganz anderes Bild: sie driften zwischen den drei Ländern viel stärker auseinander und zeigen von Süd nach Nord eine zunehmende Zurückhaltung bei der Vergabe positiver Urteile. Nur in Italien produzieren SCS und Stufenskala ähnliche Bewertungsprofile. In Deutschland scheuen die Verbraucher die Abgabe von 'Sehr gut'-Urteilen und bevorzugen stattdessen den zweitbesten Skalenpunkt (‚Gut'). Noch deutlicher ist diese Neigung in Schweden ausgeprägt. Schwedische Verbraucher ziehen sich bei ihren Beurteilungen sogar auf den Skalenmittelpunkt (‚Teils/teils') zurück. Das Resultat sind deutliche Unterschiede zwischen den Ländern.

Schaubild 4: Internationale Anwendung der SCS

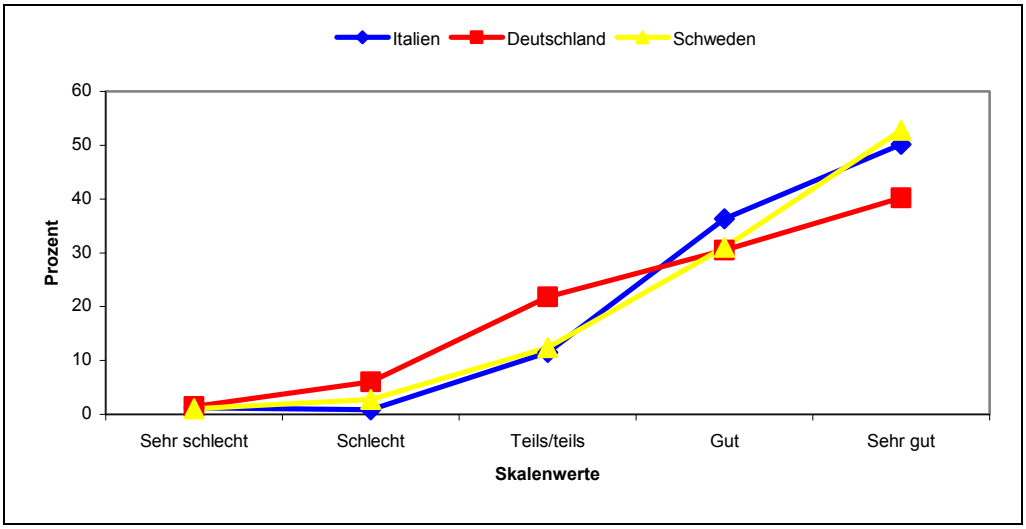
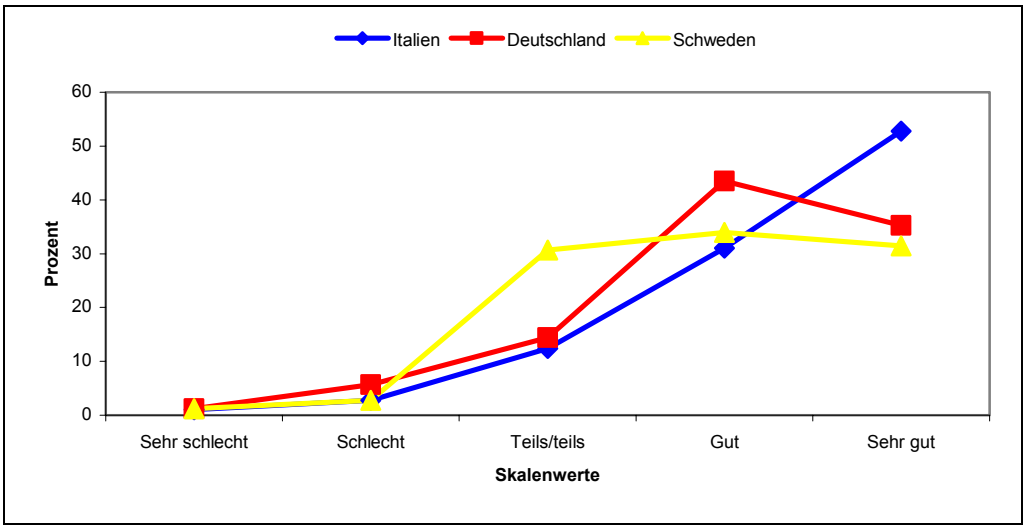


Schaubild 5: Internationale Anwendung der 5er-Stufenskala



Wenn zwei Instrumente bei der Messung desselben Sachverhalts zu unterschiedlichen Ergebnissen kommen, muss man sich fragen, welches Instrument fehlerhaft misst. Dass Stufenskalen zu einer Unterschätzung positiver Urteile führen und negative Urteile ungenügend differenzieren, ist oft vermutet worden. Hier hat sich beides bestätigt, die schlechtere Differenzierung im negativen Bereich schon in Schaubild 1.

Der Vergleich beider Schaubilder macht sichtbar, wie stark landestypisches Antwortverhalten Ergebnisse auf Stufenskalen verzerren kann. Die SCS hingegen zeigt diese Verzerrungen nicht. Aufgrund der Stufenskalen-Ergebnisse käme man zu dem Schluss, dass das getestete Premiumprodukt in Italien am besten, in Deutschland am zweitbesten, und in Schweden am schlechtesten abschneidet. Dass dies eine Fehlinterpretation ist, zeigen die SCS-Ergebnisse. Die Beurteilung des Premiumprodukts ist in allen drei Ländern tatsächlich sehr ähnlich, wenn Verzerrungen durch landestypisches Antwortverhalten aus-

geschaltet werden. Dass das Produkt wird in Italien und Schweden sehr ähnlich bewertet, in Deutschland etwas schlechter. Und das ist in der Tat ein anderes Bild als das Stufenskalen-Ergebnis, das sogar ausgesprochen nachdenklich macht: wie oft haben wohl in der Vergangenheit Skaleneffekte Unterschiede zwischen Ländern produziert, die so in Wirklichkeit gar nicht existieren?

Fazit

Dieser Beitrag schlägt vor, Werbewirkung sensibler zu messen und zeigt auch einen Weg dazu. Dieser Weg besteht darin, die Markenaffinität von Verbrauchern als psychologisches Kontinuum zu begreifen und sie vor und nach dem Werbeeinfluss auch auf einem Kontinuum zu messen. Dafür hat ISM GLOBAL DYNAMICS eigens ein neues Messinstrument entwickelt, die Stimulus Comparison Scale (SCS), das eigentlich gar keine Skala ist, sondern eine begrenzte kontinuierliche Strecke. Dieses neue Messinstrument liefert tatsächlich Daten höherer Qualität, erklärt mehr Varianz und wird auch zwischen Verbrauchern verschiedener Länder nicht so unterschiedlich verstanden und gehandhabt wie traditionelle Stufenskalen.

Abdruck mit Genehmigung des Deutschen Fachverlags, planung & analyse, Mainzer Landstraße 251, 60326 Frankfurt am Main, Tel.: 069-7595-2019, Fax 069-7595-2017, redaktion@planung-analyse.de, www.planung-analyse.de