„Recognition“ – Reconsidered[1]

Bei der Messung der Anzeigen-Beachtung erscheint die Haltung der Werbeforscher als leicht schizophren: Einerseits wird die von Dr. Daniel Starch vor 80 Jahren entwickelte „Recognition“ - Methode weitgehend mit Misstrauen betrachtet, andererseits wird sie weltweit und häufig eingesetzt.

Ich habe dergleichen Untersuchungen vor rd. 50 Jahren in Deutschland eingeführt, 1961 die Methode mit Dr. Daniel Starch, Prof. Darrell Lucas, A. Edward Miller (LIFE), Dr. Alfred Politz und anderen amerikanischen Forschern erörtert. In der nachfolgenden Überprüfung anhand sehr vieler Quellen komme ich zu einer Rehabilitation von „Recognition“.

Eine lange Geschichte

In den 80er Jahren hat der bekannte amerikanische Marktforscher und Kommunikationspsychologe Dr. Herbert Krugman die Werbeforscher dreimal aufgefordert, den Aufmerksamkeitswert der Werbung zu ermitteln (1985/86/88).

Wie anderen Marktforschern war auch ihm die Tatsache bekannt, dass in den 60er Jahren die „Recognition“-Methode zur Ermittlung der Beachtung in grundsätzlicher Weise kritisiert worden war; Lucas (1960), Appel & Blum (1961) und Wells (1964) legten hierfür die Grundlagen. Deren Kritik hat viele Forscher in aller Welt veranlasst, die Methode nur mit Bedenken anzuwenden, – wenn überhaupt.

Nach Krugmans Meinung hatte die grundlegende Untersuchung der ”Advertising Research Foundation” zu den ”Printed Advertising Rating Methods” (PARM, 1956) gezeigt, dass die Recognition-Methode in der von Starch praktizierten Art nicht so unzulänglich war, wie die Kritiker meinten. Dazu Krugman: „Obwohl einige der hauptsächlichen Ergebnisse ganz klar waren, ist nicht klar, wie es die Branche vermocht hat, diese in den nachfolgenden Jahren mehrfach zu ignorieren. Ich sage 'mehrfach', weil die Befunde von Zeit zu Zeit aufgeregt wiederentdeckt wurden, und dann entweder erneut akzeptiert oder ignoriert werden mussten. (...) es ist überfällig, die PARM-Studie von vor 30 Jahren einer ernsthaften Bewertung zu unterziehen (...) wir hoffen (...) wir werden dies alles in 10 Jahren nicht erneut tun müssen.“ (1985)

Die zehn Jahre waren bei der ursprünglichen Abfassung dieser Analyse und Veröffentlichung um; und wer die Fachliteratur der Jahre durchsieht, könnte den Eindruck gewinnen, seine Ermahnung sei gehört worden: In einer beträchtlichen Anzahl von Aufsätzen wurde das Thema direkt oder indirekt behandelt. Jedoch zeigen die nachfolgenden Kurzbeschreibungen einer Auswahl von amerikanischen Beiträgen, dass die alten irrigen Vorstellungen vorherrschen und/ oder relevantes neues Wissen nicht wahrgenommen wurde. Deshalb hat in der Messung der Beachtung von Werbung kein wirklicher Fortschritt stattgefunden.

[1] Erweiterte Fassung des im Mai 1995 in „Marketing + Reseach Today“ erschienenen Artikels

Nachfolgend einige Beispiele aus den USA von A bis Z.

Das Bruzzone-Institut hat seit geraumer Zeit Recognition-Messungen von Fernsehspots auf schriftlich-postalischem Wege durchgeführt. Aaker & Bruzzone berichteten (1981) über die Beurteilung von Fernsehspots, die zur Hauptsendezeit gesendet worden waren. Dies geschah aufgrund von 10 Adjektiven, die 4 Faktoren ergaben; diese beschrieben die wahrgenommenen TV-Spots ziemlich gut. Recognition-Ergebnisse wurden nicht gezeigt, die Methode wurde nicht erörtert.

Bagozzi & Silk haben (1983/88) eine gute Darstellung der PARM-Studie, der Geschichte von Recall- und Recognition-Messungen, relevanter Theorien und Befunde vorgelegt. Sie akzeptierten die Kritik von Lucas et al. und führten eine Reanalyse der PARM-Daten durch. Diese Analyse hat das Wissen um Recognition und Recall nicht erweitert. Howard & Sawyer (1988) sowie Finn (1992) haben kritische Anmerkungen dazu veröffentlicht.

Bogart & Tolley (1988) akzeptierten die grundsätzliche Kritik an der Recognition-Methode, was sie veranlasste, andere Messwerte zu verwenden:

- Fragen zu Produkt-Interesse, Marken-Wahl sowie Interesse an den Anzeigen
- Beobachtung des Leseverhaltens
- Messung von Gehirnströmen
- Blickregistrierung per Kamera.

Zwei ihrer lesenswerten Ergebnisse sind für diesen Beitrag relevant: Die Wichtigkeit der Zielgruppen sowie die These, dass es einen „automatischen Selektionsprozess" gäbe, der die Aufmerksamkeit der Leser auf jene Informationen lenkt, die für sie eine Bedeutung haben.

Ein Bericht von Du Plessis (1994) über Recognition vs. Recall ist ziemlich dilettantisch. Zwar wiederholt er die Kritik an Recognition nicht, geht jedoch nicht über die Feststellung hinaus, dass beides Gedächtnismessungen seien.

Finn unterzieht (1988) Starch'sche Recognition-Ergebnisse einer Analyse, indem er sie mit theoretischen Konstrukten in Verbindung bringt: Attention („beachtet"), Comprehension („seen-associated"), Elaboration („read some/most"). Dies begründet er mit Ergebnissen, die ausreichend sein mögen oder auch nicht. Er akzeptiert die Starch-Methode; seine Befunde führen nicht über das hinaus, was Starch und andere schon berichtet hatten.

Lautmann & Hsieh untersuchen (1993), ob und ggf. wie gut unterschiedliche kreative Werbemittelgestaltungen in der Lage sind, den „guten Geschmack" von Lebensmitteln vermitteln zu können. Sie diskutieren kurz Recognition als ersten Messwert der Werbeforschung. Aber sie verwenden keine Beachtungsmessung im Zusammenhang mit ihren Recall-Messungen.

Singh hat mit verschiedenen Mitarbeitern mehrfach über Recognition gearbeitet (1983/85/88). Seine Darstellung der Geschichte und der existierenden Literatur ist ge-

nerell sehr gut und nützlich. Er und seine Mitarbeiter(innen) akzeptierten einige der Kritikpunkte an der Starch-Methode. Folgerichtig haben sie Labor-Experimente mit erzwungener Beachtung und nachfolgender Recognition-Messung durchgeführt. Obwohl deren Ansatz von der Realität weiter entfernt ist als die Starch-Methode, sind einige ihrer Ergebnisse sehr wichtig:

- „Falsche“ („distractor“) Anzeigen seien problematisch (1985);
- Recognition-Werte sinken über die Zeit hinweg ab (1988);
- Recognition differenziert zwischen Anzeigen empfindlicher und stärker als Recall, und sei als Messwert so gut wie letzterer. (1988)

Tolley liefert (1993) einen umfassenden Überblick über die Zeitungsforschung in den USA; dabei erwähnt er auch Recognition-Daten für mehr als 10.000 Anzeigen, die in der Zeitspanne von 1939 bis 1954 ermittelt worden waren. Er weist ferner auf die Probleme hin, die sich bei der Ermittlung der Anzeigen-Beachtung ergeben. Den Hauptteil seines Aufsatzes widmete er der Messung der Wirkung von Anzeigen auf das Verkaufsergebnis, – wobei er ausdrücklich die Beachtungsmessung ausklammert.

Walker & von Gonten (1989) stellen ihrem Bericht über Recall-Messungen den Hinweis voran, dass man selbstverständlich zuerst ermitteln müsse, wie viele Leute den Anzeigen überhaupt Aufmerksamkeit geschenkt haben. Sie nennen ihre eigene Aufmerksamkeits-Messung ein „schwaches“ Recognition-Verfahren, – denn sie gibt den Befragten nur wenige, verbale Gedächtnis-Hilfen. Sie erkennen die Notwendigkeit, Recognition und Recall in Beziehung zu setzen, aber tun dies, indem sie schlicht die Recall-Werte als Prozentwerte bezogen auf Recognition ausweisen (mit der Bezeichnung „Linkage“ also „Verbindung“), was unsinnig ist.

Zinkhan & Gelb akzeptieren (1986) die grundsätzliche Kritik an der Starch-Methode. Sie betrachten deren Beachtungswerte als Indikatoren für Einstellungen; dabei gehen sie so weit, sie als Indikatoren der Werbewirksamkeit in Betracht zu ziehen. In ihrem Forschungsprojekt verwenden sie Messwerte für die Einstellung zu Marken und für Kaufbereitschaft. Sie finden heraus, dass Starch-Werte mit Einstellungen nicht sehr hoch korrelieren; und während Einstellungen die Kaufbereitschaft einigermassen gut vorhersagen, tragen die Starch-Werte kaum noch mehr dazu bei.

Diese kurzen Beschreibungen beschränken sich auf jene Punkte, die ich als relevant für den Zweck dieses Aufsatzes betrachte. Daraus folgt, dass sie zu kurz und einseitig sind, um den Beiträgen volle Gerechtigkeit zuteil werden zu lassen; man muss sie gründlich studieren, um ihre Bedeutung zu erkennen.

Im Gegensatz zu diesen und anderen, ähnlichen Analysen hat meine Reanalyse einen radikalen Charakter: Die Grundlagen der Kritik an der Recognition-Methode werden infrage gestellt!

Der Zweck dieses Aufsatzes ist es, Befunde und theoretische Betrachtungen vorzulegen, die nach meiner Meinung ausreichende Aussagekraft haben, um die bestehenden irrtümlichen Vorstellungen von der Recognition-Methode zu korrigieren.

Die nachfolgende Analyse ist begrenzt:

- Der Schwerpunkt liegt bei „Recognition“,
- bei Anzeigen,
- und in der Vergangenheit.

Jedoch wird die Vergangenheit vom gegenwärtigen Wissensstand her betrachtet.

Sofern diese Bemühungen befriedigende Resultate liefern, hätten wir eine Basis gefunden, um die Forschung auf Recall und andere Messwerte auszuweiten, sowie auf Werbung im Fernsehen und in anderen Medien; ferner können wir sie zu heutigen Projekten der Werbeforschung in Beziehung setzen.

Beachtung erforderlich

Die Ermahnung, den Aufmerksamkeitswert von Anzeigen zu ermitteln, beruht selbstverständlich darauf, dass die Beachtung eine notwendige, jedoch nicht ausreichende Voraussetzung der Wirksamkeit von Werbung ist. Obwohl dies eine Binsenwahrheit ist, muss sie doch ab und an wiederholt werden.

Die Frage, wie „Beachtung“ genau zu definieren ist, bleibt offen. Wir scheinen nicht viel weiter als Ebbinghaus zu sein, der vor über 100 Jahren von der „Wirkung des dunkeln Vorganges“ sprach, „den wir Aufmerksamkeit nennen“ (1880/1983, S.33).

Man kann m. E. nur sagen, dass anscheinend ein Kontinuum besteht, das vom peripheren Sehen über „Hinblicken aber nicht Sehen“ (”Looking without seeing“, Mackworth, von Thomas, 1968 zitiert) zu unterschiedlich grosser „Aufmerksamkeit“ führt. Die Beachtung befindet sich in letzterem Bereich.

Es muss überdies bedacht werden: Was durch irgendeine Recognition-Methode gemessen wird, ist offenkundig nicht die ursprüngliche Beachtung, sondern eine Art von Reflektion derselben. Die Beachtung selbst kann nicht direkt gemessen werden, ausser in einer Labor-Situation (mit deren spezifischen Schwächen).

Auch für die Ermittlung der Anzeigenbeachtung durch Recognition, oder Erinnerung durch Recall gilt die erkenntnistheoretische Feststellung: Was die Messung aufzeigt, ist als Projektion der (unbekannten) Wirklichkeit auf eine Leinwand zu betrachten (Schaefer 1989). Später entdeckte ich, dass Ekeland genau das gesagt hat (1984). Heute würde ich hinzufügen, dass die „Leinwand“ schräg stehen oder gewölbt sein kann.

Geschichte und Verwirrung

Die Geschichte der Verwendung irgendwelcher Recognition-Verfahren scheint auf Strong (1912) zurückzugehen (s. Starch 1966, König 1924/1979). Dr. Daniel Starch

hat die mit seinem Namen verbundene Methode 1922 entwickelt. Etwas später hat George Gallup sie für die redaktionelle Forschung adaptiert, und darüber seine Dissertation geschrieben (1928/86). Im Jahr 1932 begann Starch seine kontinuierlichen Umfragen zur Ermittlung der Anzeigen-Beachtung bei Leserschaften zahlreicher Publikationen.

In seinem exzellenten Buch schrieb Copland (1960) ein exzellentes Kapitel über das Thema „Konfusion". Er beschreibt und erörtert zahlreiche Tests, die anzeigen, dass Befragte Fehler machen und schnell verwirrt sind, wenn sie Auskunft über ihre Anzeigenbeachtung geben sollen. Seine älteste Quelle ist H. F. Adams (1917), der auf Experimente von Scott, Starch und Strong hinweist, in denen die vorher bestehende Bekanntheit der Testanzeigen die Ergebnisse beeinflusst haben könne. Adams erwähnt dann (leider ohne Details) eine „ingeniöse Korrektur der Bekanntheit", die von Starch erfunden wurde, – die seiner Meinung nach allerdings zu einer Überkorrektur der Resultate geführt haben könnte.

Im Jahr 1937 erfand Darrel Lucas eine Methode zur Korrektur der Konfusion (1937/38/40), die danach bei Anzeigentests und in mindestens einer Untersuchung von Verkehrsmittel-Werbung verwendet wurde (Politz 1944). Sie arbeitet mit zwei strukturgleichen Stichproben: In der einen Stichprobe wird die „Beachtung" von Anzeigen ermittelt, bevor sie veröffentlicht wurden, – die Befragten sie also nicht gesehen haben konnten. Doch für gewöhnlich erhält man trotzdem gewisse Prozente an „Beachtung". Die andere Stichprobe wird nach Veröffentlichung interviewt. Die Ergebnisse aus beiden Stichproben werden dann folgendermassen in Beziehung gesetzt:

$$\frac{\text{Post- minus Pre-Ergebnisse}}{\text{100 minus Pre-Ergebnisse}} = \text{korrigierte Ergebnisse}$$

Es ist anzuerkennen: dieses Konzept ist plausibel. Aber bei erneutem Nachdenken erscheint dergleichen Ergebniskorrektur unbefriedigend, da die Irrtümer wahrscheinlich gleichzeitig für verschiedene Anzeigen und für unterschiedliche Personen variieren, – sicherlich nicht eindimensional und querbeet.

Man kann in Copland's Buch und dem früheren von Lucas & Britt (1950) lesen, dass Starch und andere Forscher plausible und schlagkräftige Einwände gegen dergleichen Versuche der Ergebnis-Korrektur erhoben haben. Das hat im Endeffekt zur allgemeinen Auffassung geführt, dass die Resultate die extra Kosten nicht wert seien. Jedoch hat die Ungewissheit angehalten, was denn Recognition und Recall messen. Überdies waren Zweifel an der Qualität der Stichproben und deren Einfluss auf die Ergebnisse aufgekommen: Sehr kleine Quota-Stichproben. Last but not least fragten sich Werbungtreibende und Werbeagenturen, was denn die mit diesen beiden Verfahren gewonnenen Ergebnisse bedeuteten.

Die „PARM“ - Untersuchung

Die Recognition-Daten vom Starch-Institut und die Recall-Daten vom Institut Gallup & Robinson wurden in grossem Umfang verwendet, hatten deshalb einen erheblichen Einfluss auf die Werbung. Jedoch wuchsen die Zweifel an den Methoden, und so kulminierte das damit verbundene Unwohlsein zu einer von der Advertising Research Foundation geleiteten Untersuchung der Anzeigenmessungen (”Printed Advertising Rating Methods”/PARM) im Jahr 1955.

Das ursprüngliche Ziel war, mit den Worten von Lucas (1960): „ ... ein nahezu perfektes Verfahren zu entwickeln, um die Leserschaft von gedruckter Werbung zu messen und die Erinnerung daran; oder eine gründlichere Bewertung als die Leserschaftsmessungen zu erfinden, sofern möglich.“

Als ein erster Schritt wurden primär die Methoden und Verfahrensweisen der Institute Starch und Gallup & Robinson überprüft (mit der Ermittlung des „Leserinteresses“ der Fa. Readex sozusagen als Zugabe, der jedoch wenig Beachtung geschenkt wurde). Die ARF verwendete die jeweilige Verfahrensweise der drei Institute. Jedoch unterschied sich ihre Untersuchung von denen der drei kommerziellen Institute in mehrfacher Hinsicht:

1.) Da den Ergebnissen misstraut wurde, die auf Quota-Stichproben basierten, entschied sich ARF für das Random-Stichproben-Verfahren.

2.) Da man befürchtete, dass die üblichen Stichproben mit wenigen hundert Fällen inadäquat seien, wurden Stichproben mit mehr als 6.000 Interviews zugrunde gelegt.

3.) Die Ermittlung von Recognition und Recall geschah in unabhängigen Stichproben.

4.) Im Vergleich zu den Vorgehensweisen der Institute Starch und Gallup & Robinson wurden viel mehr Fragen gestellt. Und es wurden zahlreiche Auswertungen vorgenommen, um Hypothesen bezüglich der Faktoren zu überprüfen, die möglicherweise die Recognition- bzw. Recall-Werte beeinflussen.

Das Medium für die zu testenden Anzeigen war die Zeitschrift »LIFE« vom 16. Mai 1955. Parallel zur ARF-Untersuchung (von Alfred Politz Research durchgeführt) liefen bei Starch sowie Gallup & Robinson deren übliche Tests mit dieser »LIFE« - Ausgabe, sodass Ergebnis-Vergleiche angestellt werden konnten.

Angesichts der Bedeutung, die den Messverfahren und deren aufwendiger Überprüfung zugemessen wurden, erscheinen drei Sachverhalte bemerkenswert:

- Erstens wurde kein Versuch unternommen, um die Ergebnisse von Recognition, Recall oder Reader Interest auf ihren Aussagewert, geschweige denn auf ihre Validität zu überprüfen.

- Zweitens wurde auch kein Versuch unternommen, irgendein „besseres“ Verfahren zur Messung von Aufmerksamkeit/Beachtung bzw. Erinnerung zu konzipieren und im Vergleich zu testen. Eine gründliche Beschäftigung mit den Readex-Daten unterblieb, vermutlich weil man nicht recht wusste, was sie bedeuteten.

- Drittens wundert es nicht, dass unter diesen Bedingungen von der ursprünglichen Idee, ein besseres Messverfahren zu entwickeln, nur dies übrig blieb, wie mir Lucas schrieb: „Sherwood Dodge beendete die Sache damit, dass er das 'Communiscope' erfand, einen Apparat, um Anzeigen blitzartig vorzuführen, um anschliessend Recognition und Recall zu ermitteln. Ich glaube, ARF hat wohl 6 'Communiscopes' gebaut, und seitdem hat man nichts mehr davon gehört.“ (1987)

PARM - Ergebnisse positiv

Die relativ zahlreichen Auszählungen der PARM-Studie und die Vergleiche mit den Ergebnissen der parallel von Starch und Gallup & Robinson durchgeführten Umfragen lieferten praktisch keine Bestätigung der Kritik:

- Für die Recognition-Werte von PARM und Starch ergaben sich ziemlich hohe Korrelationen: r = .92 für Frauen und r = .86 für Männer.

- Die Korrelationen der PARM-Werte für Recognition mit denen von G&R lagen niedriger: r = .82, bzw. r = .61. Die Autoren der Analyse meinten, dass die niedrigere Übereinstimmung zumindest zum Teil darauf zurückzuführen sei, dass die spezielle Verfahrensweise von G&R beim Kategorisieren der Recall-Antworten von einem anderen Institut (d.h. Politz) nur begrenzt nachvollziehbar war. Das generelle Problem besteht darin, dass die Auswerter der Antworten jeweils anhand von vorgegebenen Kriterien entscheiden musste, ob tatsächlich eine Erinnerung an die betreffende Anzeige stattgefunden hat. Die Kriterien sind im hinteren Teil dieses Beitrages zitiert.

- Es gab keinen Befund, der die kleinen Quota-Stichproben der beiden Institute in Zweifel gezogen hätte.

Nur ein Ergebnis hat praktische Konsequenzen gehabt: Die PARM-Studie zeigte einen klaren Reihenfolge-Effekt, und zwar fielen sowohl Recognition- als auch Recall-Werte vom ersten zum letzten Sechstel des Interviews ab. Starch hat daraufhin bei seinen Tests die Startpunkte in den Testheften rotiert; G&R hatte bereits vorher eine Random-Reihenfolge für seine Recall-Fragen benutzt.

Somit waren die Überprüfungen der Methoden und Verfahrensweisen beruhigend und hätten deren Verwendung unterstützen können. Aber nachdem Darrell Lucas seine Position als Technischer Direktor bei der ARF verlassen hatte und Professor an der New York-University geworden war, stellte er eine ernsthafte generelle Kritik an Recognition vor (1960). Dieser Aufsatz markiert den Zeitpunkt, von dem an diese Methode als Messverfahren für „Beachtung“ in allgemeinen Verruf geriet.

Dieser schlechte Ruf der Recognition-Methode hat bis zur Gegenwart überlebt, obwohl es schon früh ziemlich positive Bewertungen gab:

- Erstens die detaillierte und meiner Meinung nach faire Erörterung von Dr. Daniel Starch in seinem Buch (1966). Die Zusammenstellung seiner eigenen und anderer Forscher Befunde zugunsten der Recognition-Methode ist eindrucksvoll. Jedoch wurde all das weitgehend unberücksichtigt gelassen oder einfach ignoriert.

- Zweitens hat der englische Forscher W. A. Twyman (1973) einen Bericht über die Messung der Doppelseiten-Kontakte und Anzeigenbeachtung verfasst.

  Er behandelte in exzellenter Weise die Themen Perzeption, Aufmerksamkeit, Gedächtnis, Recognition und Recall. Diese Arbeit hätte die wesentlichsten Vorurteile gegenüber Recognition abbauen können und sollen. Das geschah jedoch nicht, – wahrscheinlich deshalb, weil die Twyman-Analyse primär von europäischen Leserschaftsforschern gelesen wurde. Noch heute ist sie geeignet, Informationsdefizite abzubauen. Überdies liest sie sich gut: Beispielsweise prägte er für die immerwährende Wiederholung der unbegründeten Kritik an Recognition den Satz: “Rediscrediting the wheel” (als Analogie zu ”rediscovering the wheel”).

Viel später, in seinem allerletzten Aufsatz (1988), beschrieb George Gallup, wie die Anfänge seiner Adaption der Recognition-Methode in den späten 20er Jahren aussahen. In seinem typischen klaren und anschaulichen Stil legt er dar, was gute und was schlechte Verfahrensweisen sind sowie die Gründe dafür. Auch dies ist eine Verteidigung richtig angelegter Recognition-Untersuchungen und verdient (erneut) gelesen zu werden.

Last not least: Zwei Forscher des Curtis-Verlages hatten lange vor der PARM-Untersuchung ein Experiment durchgeführt, dessen Ergebnisse speziell die Anwendung der Recognition-Methode für die redaktionelle Forschung unterstützten. (Ludeke & Inglis, 1942). Den interessierten Lesern wird geraten, mindestens die grundlegenden Bücher und Aufsätze zu studieren, die hier aufgeführt werden, um die Sachverhalte im Zusammenhang zu erkennen.

Die hauptsächliche Kritik

Soweit ich es beurteilen kann, haben nacheinander Lucas (1960), Appel & Blum (1961) sowie Wells (1964) die hauptsächlichen Kritikpunkte an der Recognition-Methode als Verfahren zur Beachtungsmessung vorgetragen. Andere Forscher haben die Aussagen wiederholt oder elaboriert; jedoch beschränke ich mich auf diese drei Beiträge aus der 60er Jahren, weil sie den Ton angaben. In Deutschland sind sie durch die seinerzeit grundlegende Berichterstattung von Dr. Koeppler (1974) bekannt geworden. Aufgrund seiner zwischenzeitlichen wissenschaftlichen Beschäftigung mit der Thematik hält er heute die Kritik für weitgehend unbegründet (1995, 199?).

Die drei hauptsächlichen Kritikpunkte waren:

1.) Recognition soll eine Gedächtnisleistung messen. Das Erinnerungsvermögen des Gedächtnisses pflegt über die Zeit abzunehmen; die Recognition-Werte müssten also entsprechend abfallen.

    In der PARM-Untersuchung haben sie (über rd. 14 Tage) das nicht getan. Deshalb ist Recognition nicht das, was man annimmt. (Lucas 1960)

2.) Es wurde experimentell nachgewiesen, dass man beträchtliche Recognition-Werte für Anzeigen erhält, die die Leser der betr. Publikation gar nicht gesehen haben können.

    Andere Faktoren, wie das Produkt-Interesse oder die Nähe zum Einkauf scheinen für die Entstehung solcher falschen Recognition-Werte eine Rolle zu spielen. (Appel & Blum, 1961)

3.) Wenn damit demonstriert wurde, dass Recognition nicht wirklich „Recognition“ im ursprünglich gemeinten Sinn ist: Was ist es dann?

    Eine entsprechende Untersuchung ergab eine hohe Korrelation mit dem „Interesse“ an den einzelnen Anzeigen. Das führt zu der Annahme, dass Recognition mehr eine Messung des „Interesses“ als der „Beachtung“ sei. (Wells, 1964)

Nachfolgend meine methodologische Auseinandersetzung mit diesen drei Kritikpunkten.

Zu 1.) Sehr schwacher Abfall

Als Grundlage für seine Vorstellung, dass Gedächtniswerte im Laufe der Zeit absinken sollten, zeigte Lucas typische Vergessenskurven. In seinem Aufsatz von 1960 gab er hierfür leider keine Quelle an, aber in ihrem Buch von 1963 beziehen sich Lucas & Britt auf eine Veröffentlichung von Hovland et al. (1953). Meine Nachforschung dort endete bei der Untersuchung von Recognition und Recall von sinnlosen Silben durch einen Nachfolger von Ebbinghaus (Luh 1922).

Ich stimme nicht ganz der grandiosen Verallgemeinerung von Prof. Ulric Neisser zu, dass das Studium sinnloser Silben „wahrscheinlich der Archetyp psychologischer Irrelevanz“ sei. (1979, S. 46) Nach meiner Kenntnis war es ein intelligenter Anfang von Ebbinghaus, dem weitere Schritte von ihm und/oder anderen Forschern hätten folgen sollen. Jedoch: Die Befunde zu sinnlosen Silben haben nur für sinnlose Silben einen Sinn, aber für die Erforschung der Realität sind sie nicht verwendbar. Bereits Hovland et al. hatten darauf hingewiesen, dass Recognition und Recall von sinnlosen Silben einerseits und sinnvolles Material andererseits zwei ganz verschiedene Dinge seien.

Hinzu kommt, dass Recall (von sinnlosen Silben oder sonstwas) nicht dasselbe wie Recognition ist. Das bedeutet, dass Ebbinghaus'sche Erinnerungs-Kurven beruhend

auf Recall irrelevant für Recognition sind. Lucas hätte sich nicht darauf beziehen dürfen.

Im übrigen können auch Recognition-Kurven absinken, – langsam, wie vom Starch-Institut gefunden wurde (1958). Aber es scheint, als habe niemand deren Ergebnisse ernst genommen. Darüber hinaus gibt es eindeutige Befunde, dass Recognition über die Zeit hinweg recht stabil bleibt: Mit einem sehr einfallsreichen Testverfahren fanden Nowak & Smith (1971), dass Recognition-Werte „mindestens für ein Jahr“ auf ihrem Niveau blieben; das ist erheblich länger als die 2 Wochen in der PARM-Studie, auf die sich Lucas bezog.

Ein anderes relevantes, wenngleich extremes Untersuchungsergebnis sollte zur Kenntnis genommen werden: Bahrick et al. haben (1975) berichtet, dass der Mensch andere Personen über einen Zeitraum von 50 Jahren korrekt wiedererkennt!

Aus allen diesen Befunden ergibt sich die Schlussfolgerung, dass die Annahme abgelehnt werden muss, die Recognition-Werte in der PARM-Studie hätten innerhalb von 14 Tagen absinken müssen. Also fehlt der Lucas'schen grundlegenden Kritik die Basis.

Zu 2.) Falsche Antworten ?

In einem Recognition-Experiment hatten Appel & Blum (1961) einer Stichprobe von Lesern der Zeitschrift »LIFE« eine Ausgabe vorgelegt, die sie gelesen hatten, – jedoch mit kunstvoll eingefügten Anzeigen, die sie also nicht gesehen haben konnten. Durchschnittsergebnis: 25% „Recognition“ für diese „falschen“ Anzeigen. Dies war die Grundlage für ihre oben zitierte Kritik an der Methode.

Dr. Morgan Neu vom Starch-Institut duplizierte dieses Experiment und erhielt 24% „Recognition“ für die falschen Anzeigen. Klarer Fall ?

Nein: In einer zweiten, strukturgleichen Stichprobe des Neu'schen Experiments wurde den Befragten gesagt, dass sich in dem vorgelegten Heft Anzeigen befanden, die in dem ursprünglich von ihnen gelesenen Heft nicht enthalten waren; sie sagten jedoch nicht, welche es sind. Die Irrtums-Quote sank auf 6%! Eine Wiederholung dieses Experiments bestätigte das Ergebnis.

Neu's Schlussfolgerung: Die falschen Appel-Blum'schen Recognition-Ergebnisse wurden durch deren falsche Versuchsanordnung produziert.

Es gibt relevante Befunde aus einem ganz anderen Bereich: Hall und Rostrum fanden (1975) dasselbe Phänomen beim Versuch der Identifikation von Tatverdächtigen durch Zeugen bei der Gegenüberstellung mit einer Reihe von Personen. Die simple Vorgehensweise liefert 28% falsche Identifikationen.

Wenn den Zeugen jedoch gesagt wird, dass möglicherweise die gesuchte Person sich garnicht unter den vorgestellten befindet, sinkt die Irrtums-Quote auf 4%! (Degen

1981) Nota bene: Die korrekte Identifikation wird durch diese Instruktion nicht verringert. Malpass & Devin (1984) zitieren gleichartige Befunde.

Den Einfluss von Instruktionen – oder deren Abwesenheit, die es den Testteilnehmern überlässt, ihren eigenen Vorstellungen oder Annahmen zu folgen, was denn von ihnen erwartet wird (s. Rosch 1978, S.43) – hat man in einer Reihe von Experimenten gefunden; Beispiele enthält die ausgezeichnete Aufsatz-Sammlung von Brown (1976).

Zusätzlich zum potentiell verfälschenden Einfluss von unzulänglichen, falschen oder gar keinen Instruktionen haben wir das Problem des Kontextes. In der akademischen Forschung war die Veränderung des dargebotenen Kontextes Bestanteil vieler Experimente. Sofern dies nicht aus Versehen geschah, war es deren Ziel, herauszufinden, ob und ggf. in welchem Mass das Ergebnis von Recognition und Recall vom Umfeld abhängt. Endel Tulving stellt als typischen Befund fest, dass eine Veränderung des Kontextes zwischen ursprünglicher Vorlage und späterem Test eine Beeinträchtigung der Recognition-Leistung zur Folge hat. (in Brown, S. 53)

Zu diesem Thema ist kürzlich ein ganzes Buch herausgekommen, mit dem schönen Titel: ”Memory in Context – Context in Memory” (Davies & Thomson, 1988). Thomson kommt zum gleichen Resultat (S. 285). Der potentielle Einfluss von Instruktionen und Kontext muss im Licht des heutigen Wissens um das Gedächtnis gesehen werden. Das wird weiter unten geschehen.

Es ist demnach so, wie es Neu bereits gesagt hatte: Die falschen Recognition-Werte, die von Experimenten wie dem von Appel & Blum geliefert wurden, sind weitgehend Artefakte der Versuchsanordnung. Sie haben keinen Beweiswert gegen die Recognition-Methode.

Das heisst, auch der zweite Kritikpunkt ist unbegründet.

Zwischenfrage: Wie funktioniert die Erinnerung?

Um zu verstehen, warum Instruktionen und Kontext die Recognition- und Recall- Werte beeinflussen können, und „falsche“ Anzeigen im normalen Zeitschriften-Umfeld Ergebnisse produzieren, ist in Betracht zu ziehen, was zum Gedächtnis bekannt ist. Das Studium von diversen Büchern über Kognition und Gedächtnis zeigt Übereinstimmung hinsichtlich zweier grundlegender Sachverhalte:

Erstens gibt es so etwas wie ein „photographisches“ Gedächtnis praktisch nicht. In der Evolution von Tieren und Menschen war die Entwicklung einer solchen Fähigkeit nicht nötig. Was gebraucht wurde, hat Ulric Neisser so formuliert: „Wir müssen die Bedeutung und nicht oberflächliche Details erinnern.“ (1982) Die relevanten biologischen Grundlagen wurden von Rupert Riedl sehr gut dargestellt (1980).

Zu dieser Thematik gibt es eine faszinierende Analyse. Nixon's Mitarbeiter John Dean hatte bei seiner Anhörung zur Watergate-Affäre im US-Senat alle Welt durch seine scheinbar „wörtliche“ Wiedergabe relevanter Diskussionen im Weißen Haus zwischen Nixon und anderen Personen beeindruckt. Wie bekannt, tauchten später Tonbänder

mit u.a. diesen Diskussionen auf. Prof. Neisser und Mitarbeiter haben diese mit Dean's Aussagen verglichen: Als eine anscheinend „perfekte“ Wiedergabe dessen, was von wem gesagt wurde, waren Dean's Aussagen voller Fehler, – aber alles Wesentliche hatte er ziemlich korrekt dargestellt.

Zweitens: Da das Gedächtnis nicht alle Details aufbewahrt (und jene Details, die gespeichert werden, nicht von photographischer Genauigkeit sind), findet bei der Erinnerung fast immer ein Prozess der Rekonstruktion statt – außer bei einfachsten Sachverhalten.

In der Literatur wird dieser Befund auf den bedeutenden britischen Psychologen Sir Frederic Bartlett zurückgeführt (1932). Erst neuerdings wurde in Deutschland darauf hingewiesen, dass der bis 1933 in Hamburg lehrende Psychologe William Stern dergleichen bereits 1914 gefunden hatte. (Hartmann 1995) Kürzlich gab es einen weiteren historischen Hinweis (Assmann 2003), dass Maurice Halbwachs 1925 auf dieser Schiene war (Halbwachs 1985).

Peter Kelvin hatte u.a. auf diesen Sachverhalt der Rekonstruktion in seiner gut geschriebenen und leicht verständlichen Darstellung von ”Advertising and Human Memory” (1962) hingewiesen. Man liest dergleichen Zitate in vielen Fachbüchern.

Wenn das so ist, dann müssen sich die Befragten offensichtlich auf ihre Erinnerungsbruchstücke, Erwartungen, Annahmen, Vorurteile sowie auf Schlussfolgerungen verlassen, um irgend etwas zusammenhängend und überzeugend zu rekonstruieren, – sei es gegenüber einem Polizisten, einem Richter, einem Interviewer, oder dem Ehepartner. An dieser Stelle spielen korrekte oder falsche Hilfen ihre nützliche oder verfälschende Rolle.

Falls die Zeugen bei der Verdächtigen-Identifizierung keine faire Instruktion oder Warnung erhalten, werden sie unterstellen, dass der/die Gesuchte sich unter den aufgereihten Personen befindet. Sie antworten dann im Sinne einer Rekonstruktion von: Hat eine(r) von diesen eine Ähnlichkeit mit der Person, die ich gesehen hatte? Sie rekonstruieren anders, als wenn sie darauf aufmerksam gemacht wurden, dass sich die gesuchte Person möglicherweise nicht unter den vorgeführten befindet.

Wenn die Befragten bei einem Recognition-Test mit einer Zeitschriften-Ausgabe konfrontiert werden, die sie bereits gelesen haben, argwöhnen sie normalerweise nicht, dass sie von den Befragern hinters Licht geführt werden. Sie gehen die Ausgabe unbefangen durch, erkennen viele redaktionelle Beiträge und Anzeigen wieder, – sowohl solche, die sie beachtet haben wie jene, die sie ignorierten; aber das meiste erscheint ihnen ziemlich gut bekannt. Sie rekonstruieren (weitgehend unbewusst) ihr Verhalten – ihre Wahrnehmung wie ihr Lesen – in etwa dieser Weise: Ich habe diese Seite aufgeschlagen, den Artikel oder diese Bilder auf der linken Seite beachtet/gelesen, deshalb habe ich wohl auch die Anzeige auf der rechten Seite gesehen. Sie (die Anzeige u/o die Marke) kommt mir bekannt vor, also werde ich ihr irgendeine Art von Aufmerksamkeit geschenkt haben.

Starch (1966) und Gallup (1986) haben darauf hingewiesen, dass es wichtig ist, die Anzeigen in ihrer vollständigen, „natürlichen“ Umgebung zu präsentieren. Alfred Politz

hat in derselben Weise für seine Methode der Identifikation von Zeitschriften-Ausgaben argumentiert: Die Stimuli für Kognition und Recognition sollen identisch sein. (LIFE 1953) Die Gedächtnisstützen, die somit für das korrekte Wiedererkennen von redaktionellen Beiträgen, Zeitschriften-Ausgaben oder Anzeigen dienlich sind, können ggf. als eine gleichfalls sehr wirksame Ursache der Falschidentifikation funktionieren.

Auf einen dritten Punkt wurde ich aufmerksam, als ich einige Aussagen in der englischen und der deutschen Fassung von Prof. John Young's Buch „Philosophie und Gehirn“ (1986/89) verglich. In zwei Fällen wurde sein englischer Begriff “Recognition” mit „Erkennen / Erkennung“ übersetzt. Das ist ein ärgerlicher Fehler, denn dadurch ging seine implizierte, korrekte und wichtige Beobachtung verloren: Wir sollten wissen, dass das meiste von dem, was wir als Erkennen betrachten, ein Wiedererkennen ist. Kein Lebewesen könnte die Welt verstehen und überleben, wenn es nicht das meiste von dem, was es sieht, hört, fühlt, schmeckt oder riecht, bereits kennen würde, so dass es identifiziert, d.h. „wieder erkannt“ wird, wenn es (wieder) auftritt.

Diese erkenntnistheoretisch wesentliche Beobachtung führt uns zur Hypothese, dass Kognition und Recognition denselben Regeln folgen und denselben Einflussfaktoren unterworfen sein sollten.

Daraus ist auch die Schlussfolgerung zu ziehen, dass es beinahe hoffnungslos sein dürfte, unter „natürlichen“ Bedingungen zwischen den beiden zu unterscheiden. Um ein Beispiel zu nennen: Die voraufgegangene Veröffentlichung der gleichen oder sehr ähnlicher Anzeigen kann (Re-)Cognition wie auch (Re-)Recognition beeinflussen. Wie könnte man das trennen?

Die nächste Frage: Kann man daraus schliessen, welcher Art die „falschen“ Recognition-Ergebnisse sind, und wie sie sich zu den „richtigen“ verhalten?

Die beiden Datensätze sollten sich ähnlich sein, jedoch auf unterschiedlichen Niveaus. Mit anderen Worten: die falschen und richtigen Ergebnisse sollten ziemlich hoch korrelieren. Die Korrelation sollte in dem Mass verringert werden, wie die falschen Anzeigen selbst wenig oder garnicht bekannt sind bzw. für wenig/gar nicht bekannte Marken werben.

Wir müssen nicht spekulieren, was herauskommt, denn Appel & Blum hatten auch diese Frage behandelt: Als Teil ihres Experiments hatten sie eine strukturgleiche Stichprobe hinzugefügt, die vor der Veröffentlichung der Testausgabe befragt wurde. (Appel 1993) Diese Befragten konnten also weder die falschen noch die richtigen Anzeigen gesehen haben, jedoch kamen für beide „Recognition“-Werte heraus. So konnten die Autoren die Korrelation zwischen falschen (vorher) und richtigen (nachher) Ergebnissen rechnen. Die Korrelation war ziemlich hoch: r =.72; d.h., ungefähr die Hälfte der Variation des einen Datensatzes konnte durch die Variation des anderen „erklärt“ werden.

Quod erat demonstrandum!

Zu 3.) Interesse, Attraktivität und persönliche Relevanz

Wells hatte (1964) Lucas' grundsätzliche Kritik, sowie die Befunde und Interpretationen von Appel & Blum akzeptiert und sich deren These angeschlossen, dass „Interesse“ ein wesentlicher Einflussfaktor für die (falsche) Recognition zu sein scheint. Er ging so weit, zu formulieren, „dass Recognition - Ergebnisse gar nicht Recognition-Ergebnisse sind.“

Er bezog sich auf die relativ hohe Korrelation zwischen den Ergebnissen der „Leserinteresse“ - Ermittlung und denen für Recognition in der PARM-Studie, insbesondere bei Frauen. Er vermutete, „dass Recognition mehr ein Ausdruck von Interesse als ein Mass für Recall“ sei.“ (Es ist nicht verständlich, warum er hier Recognition mit Recall gleichsetzt.)

Dazu hat er über die Ergebnisse einer Untersuchung berichtet, in der Attraktivität und Bedeutsamkeit der Anzeigen mit Recognition und Recall in Beziehung gesetzt wurden: Recognition korrelierte höher mit Attraktivität, Recall höher mit Bedeutsamkeit.

Wenn aber die zwei grundlegenden Annahmen unbegründet sind, wie oben dargestellt, dann verliert die ganze Wells'sche Argumentation ihre Basis.

Jedoch sind die Daten von Wells weder falsch noch irrelevant; im Gegenteil, sie liefern gültige und wertvolle Erklärungen für hohe, mittlere und niedrige Recognition- und Recall-Werte. Wir sollten mit der Frage anfangen, *warum* jemand eine Anzeige beachtet. Wells war auf der richtigen Spur, sowohl in seiner PARM-Analyse als auch in seinen nachfolgenden Projekten zur Erforschung relevanter Eigenschaften von Werbemitteln (1964/ 1971): Ob eine Anzeige „interessant“ ist (oder nicht), hat einen grossen Einfluss.

Jedoch ist das „Interesse“ ein nicht gerade sehr klarer und eindeutiger Begriff: Interesse am Produkt, an der Werbebotschaft, am Modell oder an was?

Eine Studie von March & Swinbourne (1974) weist nach, dass die wesentlichste Komponente von „Interesse“ das „Interesse an der Produkt-Botschaft“ ist. Diese wiederum wird hauptsächlich durch die „persönliche Relevanz“ erklärt, wie sie in dem von Wells et al. entwickelten Test (1971) gemessen wird.

Gehen wir nun zur „Attraktivität“ der Anzeige: Dieser Begriff scheint ziemlich klar zu sein. Jedoch haben Biel & Bridgwater (1990) demonstriert, dass mehr dahinter steckt, als es auf den ersten Blick scheint: Wenn ein Tier(*chen*) in dem Werbemittel dargestellt wird, dann hat die Attraktivät eine spezielle Bedeutung (die man wohl nicht erklären muss). Schliesst man Werbemittel mit Hündchen, Kätzchen etc. von der Untersuchung aus, dann ist der hauptsächliche Einflussfaktor für die Attraktivitätsergebnisse das Syndrom Relevanz/ Wichtigkeit/ Bedeutsamkeit.

Das sollte uns an den „klassischen“ Befund von Alfred Politz erinnern, dass die subjektive Wichtigkeit der Werbebotschaft auf die Werbewirksamkeit den grössten Einfluss hat (1955).

Wenden wir uns nun dem Wells'schen Befund zu, dass die Bedeutsamkeit eine höhere Korrelation mit Recall als mit Recognition hat.

Meiner Meinung nach verstehen wir derzeit noch nicht so recht, was Anzeigen- "Recall" eigentlich bedeutet, obwohl Dubow (1994) einen vielversprechenden neuen Anfang gemacht hat. Jedoch scheint es mir sicher zu sein, dass eine Art selektiver Prozess stattgefunden hat, in einem Ablauf von

→ Sehen
→ beachten
→ Aufmerksamkeit schenken
→ im Gedächtnis speichern
→ erinnern.

Aber dieses Modell sagt nichts zur Bedeutung von Recall für die Werbung.

Warum erinnert man sich an gewisse Anzeigen und vergisst andere (sehr) schnell ? Wiederum sollte deren Bedeutsamkeit in dem Selektionprozess bis zur Erinnerung eine grosse Rolle spielen, – plausiblerweise mehr als für das ursprüngliche Sehen/ Beachten.

Und genau das hatte Wells herausgefunden. Andere Forscher, die Recall - Werte analysiert haben, dürften bestätigende Befunde vorliegen haben (z.B. G & R 1990, S.2).

Damit haben wir festgestellt, dass Recognition nicht Interesse ist. Vielmehr dürfte das Interesse ein kausaler Einflussfaktor für Recognition sein. Kann diese Vorstellung begründet werden?

Jene Forscher, die per Augenkamera Blickregistrierung durchführen, können uns hier helfen. Thomas (1968) hatte auf das Phänomen "looking without seeing" hingewiesen, das von Mackworth beobachtet worden war. Dies beschreibt einen Unterschied zwischen zielgerichtetem und ziellosem Wandern des Blicks und dem, was man Aufmerksamkeit nennt. Thomas erläutert dies: „Es ist verblüffend, wenn man einen Film der eigenen Augenbewegungen ansieht. Der zeigt hunderte von Fixationspunkten des Blicks bezogen auf Einzelheiten, von denen man nicht die geringste Erinnerung hat."

Als Erklärung benutzte er ein Gleichnis: „Teile des Gehirns funktionieren anscheinend wie eine Sekretärin, die Routineangelegenheiten ohne Rücksprache mit ihrem Arbeitgeber erledigt und ihn auf wichtige Punkte in einlaufenden Briefen aufmerksam macht, – die aber gelegentlich Fehler macht." Man beachte das Wort „wichtig". Noton & Stark fassten zusammen: „Die Augenmuskeln, kontrolliert vom Gehirn, lenken den Blick auf Punkte, die von Interesse sind." (1971)

Krugman war (wie so oft) auch schon auf dieser Fährte. (1971) Er zitiert den berühmten Psychologen William James, der bereits vor 115 Jahren sagte: „Die freiwillige Aufmerksamkeit ist immer abgeleitet; wir bemühen uns niemals, einem Objekt Aufmerksamkeit zu schenken, außer wenn diese Bemühung irgendeinem Interesse dient." (1890)

Um diese historische Aussage um 10 Jahre zu übertreffen, darf Ebbinghaus zitiert werden, der in seinem wiedergefundenen Manuskript (1880/1983, S.70) über die Einflüsse „ ... durch das ausserordentlich verschiedene Interesse – im weitesten Sinne – welches verschiedene Vorstellungen verschieden fest haften macht“ schrieb. Das war der Grund dafür, dass er sinnlose Silben als Gegenstand seiner Versuche wählte, bei denen das „Interesse“ keine Rolle spielen konnte.

Auf diesem Wege sind wir im vollen Kreisbogen auf das Syndrom Interesse/ Relevanz/ Wichtigkeit gekommen, das ein kausaler Einflussfaktor für sowohl (periphere) Wahrnehmung, als auch für Aufmerksamkeit, Kognition, Recognition und Recall ist.

Kritik widerlegt

Schlussfolgerung: Die drei hauptsächlichen Kritikpunkte an der auf Anzeigen bezogenen Recognition-Methode sind widerlegt.

Aufgrund der hier zutage geförderten Ergebnisse und Erkenntnisse können wir einen neuen Anfang machen. Die PARM-Untersuchung soll als Ausgangspunkt dienen: Zunächst werden wir etwas mehr an Wissen zur Bedeutung von Recognition und Recall gewinnen; zweitens werden wir uns mit der Wichtigkeit der Zielgruppen befassen.

Recognition versus Recall

Die Erörterung der Bedeutung und des Messens von Recall und Recognition durch eine Reihe gelehrter Autoren in der Sammlung von Brown (1976) scheint auf zwei Schulen hinzudeuten: Die erste ist der Meinung, dass Recall und Recognition ein Kontinuum darstellen; die zweite meint, dass diese beiden Messwerte verschiedene Aspekte des Gedächtnisses widerspiegeln. Jedoch war es mir nicht möglich, aus dieser (oder einer anderen) akademischen Quelle Informationen zu extrahieren, die für die Werbeforschung relevant erscheinen.

Dieser Misserfolg ist gleichsam ein Echo dessen, was Howard & Sawyer (1988) in ihrer Kritik an der Analyse von Bagozzi & Silk (1983) vortrugen: Sie stellen die Annahme infrage, man könne aufgrund aggregierter Prozentzahlen aus unabhängigen Stichproben sinnvolle Rückschlüsse auf individuelle Gedächtnisprozesse ziehen: „Die PARM-Daten stellen Eigenschaften von Anzeigen und nicht von Individuen dar.“

Ihre letzte Feststellung sollte nicht als die Beschreibung einer bedauerlichen Einschränkung angesehen werden. Unsere Aufgabe ist nicht das Testen von Personen, sondern von Produkten, Packungen, Parfüms, Publikationen, – und hier: von Print-Werbung. Das bedeutet, dass wir an den Eigenschaften von Anzeigen interessiert sind sowie an den Gründen für hohe, mittlere und niedrige Messwerte.

Deshalb sollten wir uns den Praktikern der Forschung zuwenden, um zu sehen, was sie zu sagen haben.

Recognition und Recall beruhen beide auf dem „Gedächtnis“; das ist ihnen gemeinsam. Die Methoden, um Recall und Recognition zu ermitteln, sind selbstverständlich unterschiedlich, aber die Messungen befinden sich auf einem Kontinuum, wie Alfred Politz erläutert hat:

> „Wenn sie mit dem Problem der Messung konfrontiert werden, müssen sich die Forscher häufig zwischen den Alternativen Recall und Recognition entscheiden. Diese Bezeichnungen, wie sie in der Praxis benutzt werden, stellen eine Übervereinfachung dar. In Wirklichkeit ist der Unterschied zwischen Recall und Recognition ein gradueller. Jeglicher Recall wird durch externe oder interne Stimuli gestützt. Je grösser die Ähnlichkeit von Stimulus und ursprünglichem, zu erinnernden Objekt ist, desto mehr neigen wir dazu, von ’aided recall’ zu sprechen. Wenn die Ähnlichkeit sich der Identität nähert, neigen wir dazu, von ’recognition’ zu sprechen.“ (LIFE, 1953, S. 8)

Dies ist eine nützliche, grundlegende Beschreibung, – aber was bedeuten die von den beiden Methoden gelieferten, unterschiedlichen Ergebnisse?

Sherwood Dodge beschrieb die Sache so: „Sie mögen mich fragen, ob diese zwei Methoden grundsätzlich dieselbe Sache beschreiben oder nicht. Einige von Ihnen haben große Zweifel, aber ich muss meine Auffassung als ein eingeschränktes 'ja' vortragen. (...) Die von den beiden Methoden produzierten Werte sind Ergebnisse von Gedächtnis-Tests mit zwei Schwierigkeitsgraden – wobei ’aided recall’ der weit schwierigere ist.“ (1956)

Aber auch „unterschiedliche Schwierigkeitsgrade“ sagen nichts darüber, was die beiden Methoden messen. Wie wir alsbald sehen werden, hat Dodge eine Analyse durchgeführt, deren Ergebnisse darauf schliessen lassen, dass die zwei Methoden (etwas) unterschiedliche Sachverhalte messen.

In seiner PARM-Analyse bot Darrell Lucas (1961, S. 11) diese Vorstellung an: „Das Komitee liess mit voller Absicht diese parallelen Tests durchführen, um die beiden kommerziellen Dienste vergleichen zu können. Der Autor meint, der Vergleich ist von minderer Bedeutung, und zwar aufgrund der Voraussetzung, dass die beiden Dienste zwei wesentlich unterschiedliche Dinge messen sollen und es wahrscheinlich tun.“ (Anm.: Mit den kommerziellen Diensten meinte er die der Institute Starch und Gallup & Robinson).

Lucas spricht nicht aus, was denn die beiden Methoden messen. An einer Stelle gibt er zu, dass Recognition-Ergebnisse Indikatoren für „Beachtung“ sein mögen, – so wie von Starch selbst bezeichnet; und das vergleicht er mit ”Exposure” von Anzeigen. (Anm.: ”Exposure” bezieht sich auf die seinerzeit von Politz durchgeführte Messung der sogen „Doppelseiten-Kontakte“ = ”Advertising Page Exposure”.) Aber er geht dann nicht weiter, etwa im Sinne der oben zitierten Beobachtung ”looking but not seeing”, die ihn möglicherweise veranlasst hätte, weniger vage zu definieren.

Bezüglich „Recall“ bietet Lucas keine Definition an, sondern formuliert schlicht: „Gallup & Robinson nehmen in Anspruch, Recall zu messen, und die Prozentwerte verhalten sich wie Recall - Prozentwerte.“ (1961, S.17)

Krugman stellte (1977/85/86) eine Neubetrachtung der (höheren und stabilen) Recognition - Werte und der (niedrigeren und schnell absinkenden) Recall - Werte an, und zwar auf der Grundlage der Gehirnforschung; diese deutet darauf hin, dass normalerweise die linke Gehirnhälfte mehr für Worte, die rechte mehr für Bilder verantwortlich ist. Er schliesst daraus: „Wir wissen, dass das Gedächtnis für Bilder und Images rechtsseitig im Gehirn extrem gut und das Vergessen gering ist, – wohingegen das Gedächtnis für Worte relativ gering und mit starkem Vergessen behaftet ist. (...) was sehr deutlich darauf hinweist, dass Recognition und Recall zwar beides Gedächtnismessungen sind, aber Messungen unterschiedlicher Gedächtnisinhalte."

Dazu: Haber (1970) und Standing (1973) haben von korrektem Wiedererkennen von tausenden von Bildern berichtet. Eine vollständige Beschreibung von linker und rechter Gehirnhälfte liefern Springer & Deutsch (1985/87). Neuere Ergebnisse findet man in einem Buch von Livingston (1989), neueste Befunde wurden von Raichle (1994) vorgestellt.

All dies sind wertvolle Hintergrund-Informationen; sie mögen uns im Laufe der nächsten Zeit und der rasant fortschreitenden neurologischen Forschung behilflich sein, die unterschiedlichen Ergebnisse unserer diversen Methoden besser zu verstehen.

Da jedweder Messwert grundsätzlich nicht für sich allein betrachtet definieren lässt, müssen in einem ersten Schritt verschiedene Messwerte miteinander in Beziehung gesetzt werden. Nur so ist herauszufinden, was sie über unterschiedliche Anzeigen sowie über sich selbst aussagen.

X & Y in der PARM - Studie

Der aufmerksame Leser wird bei der Zwischenüberschrift die Anspielung auf den zitierten Aufsatz von Lucas (1960) wie auch auf die Regressionsanalyse bemerkt haben. Sherwood Dodge und Jerome D. Greene (damals bei Politz) hatten Regressionsanalysen durchgeführt (Dodge 1956). „Zahlreiche" schrieb mir Greene (1990); auffindbar waren ganz wenige. Sie weisen aus, dass Recognition und Recall korrelieren: Bei den Frauen r = .7, also 49% erklärte Varianz.

Sie haben die Resultate ihrer Regressionsanalysen aufgezeichnet, um zu sehen, für welche Arten von Anzeigen sich höhere bzw. niedrigere Messwerte als die Erwartungswerte ergaben. Folgende Arten von Anzeigen lagen mit ihren Recall - Werten deutlich über den Erwartungswerten:

1.) Einfachere, leichter zu lernende Anzeigen.
2.) Anzeigen, die den vorgegebenen Gedächtnisstützen näher standen (Karten mit Produkt- und Markeninformationen).
3.) Anzeigen für gut bekannte Marken.
4.) Anzeigen in Produkt-Gruppen, für die ein höheres Interesse existierte.

Anzeigen mit unter den Erwartungswerten liegendem Recall spiegeln diese Ergebnisse wider, mit einer Ausnahme: Sofern die betreffende Ausgabe von »LIFE« Anzeigen für andere Marken aus der jeweiligen Produkt-Gruppe enthielt, tendierten die Recall-

Werte dazu, unterhalb der Erwartung zu liegen. Ob das die „Wirklichkeit“ reflektiert oder ein Artefakt der Testanlage ist, wüsste man gerne.

Dodge war der Meinung, dass viele weitere derartige Befunde aus der Untersuchung gewonnen werden könnten und schlug vor, dafür $ 20.000 einzusammeln. Sein Plan ist, soweit mir bekannt, nie realisiert worden. Postscript AD 2005: Wie schade !

Bedeutung der Zielgruppen

Als ich vor rd. 20 Jahren ernsthaft damit begann, die PARM-Untersuchung und ihre Ergebnisse erneut zu studieren, dachte ich, dass die Daten benutzt werden könnten, um mehr über die Beziehung zwischen Recognition und Recall herauszufinden. Allerdings sollte das auf einer etwas anderen und korrekteren Grundlage erfolgen, als das bisher geschehen war.

PARM zeigte, dass das Merkmal „Interessenten“ (”prospects”) einer Produktgruppe von erheblicher Bedeutung für die Ergebnisse der Messungen war: Bei den Recognition-Werten betrug der Unterschied zwischen Interessenten und Nichtinteressenten im Schnitt relativ rd. 50%, bei Recall 100%. Dergleichen Befunde waren selbstverständlich schon in vielen anderen Untersuchungen beobachtet worden. Sehr frühe Quellen und Ergebnisse stellte Copland vor. (1958/ 60) (Anm.: Es wird auf die Problematik der „relativen“ Prozentzahlen im Kapitel . . . hingewiesen!)

Trotz dieser Befunde werden die Ergebnisse von Recognition und Recall auch noch heutzutage auf der Basis von „allen Befragten“ ausgewiesen. William Wells – der sich im Brief zu meinem englischen Entwurf dieser Analyse als Befürworter von Recognition bezeichnete (1993) – hält es für einen wesentlichen Nachteil, dass „gegenwärtige standardisierte Recognition-Meßsysteme nicht die Interessenten von den Nichtinteressenten separieren.“ Es wird unterstellt, dass praktisch jedermann potentieller Kunde sei (z.B. Starch 1966, S.111). Bereits Copland hatte in seinem Buch Beweise dafür gebracht, dass dies ein Fehler ist, wobei er unter anderen Hans Zeisel (1951) zitierte.

Es gibt offensichtliche Fälle wie Zigaretten-Anzeigen, die kaum in der Lage sein dürften, aus Nichtrauchern Raucher der betr. (oder irgendeiner anderen) Marke zu machen. Nichtraucher mögen sich auch Zigaretten-Anzeigen ansehen, aber in einem deutlich geringeren Ausmaß und aus anderen Gründen als Raucher. Das oben erwähnte 4. Ergebnis von Dodge passt dazu.

Wesley Rosberg hatte bereits 1956 darauf hingewiesen, dass sie ihre Analysen auf die tatsächlichen Interessenten für die jeweiligen industriellen Güter und Dienstleistungen beschränken mussten, um herausfinden zu können, welche Faktoren die Beachtung industrieller Anzeigen beeinflussen.

Auf eine eigene Untersuchung ist hier hinzuweisen, die in diesem Buch im Kapitel ... beschrieben wird. Sie bezieht sich auf Anzeigen des Lebensmittel-Einzelhandels (1984). Sie weist nach, dass die gewohnheitsmäßige Anzeigen-Beachtung sehr stark davon abhängt, ob die befragte Person häufige, gelegentliche oder Nicht- Kundin des betr. Anbieters ist.

Anders betrachtet: Die auf alle Personen bezogenen Durchschnittswerte der Beachtung korrelieren sehr hoch mit den Umfängen der Käuferschaften der einzelnen Handelsunternehmen. Legt man einer Analyse schlicht nur solche Durchschnittswerte zugrunde, landet man schnell bei den üblichen Ergebnissen: Grosse u/o farbige Anzeigen erzielen höhere Werte. Dabei dürfte deren Verwendung in der Regel von der Finanzkraft der Anbieter, diese vom Umfang der jeweiligen Käuferschaft abhängen.

Es ist aufgrund derartiger Sachverhalte kein Wunder, was Willke (1993) herausfand: dass die vom amerikanischen Institut „ASI“ ermittelten Normwerte für die einzelnen Produktkategorien positiv mit der Grösse der Produkt-Kategorie korrelieren, – wo doch deren „Normen“ auf Jedermann beruhen.

Um dergleichen Probleme bei der Reanalyse der PARM-Ergebnisse zu vermeiden, versuchte ich, die vorhandenen „Interessenten“-Messwerte zu verwenden. Lucas hatte sie als „grobe Klassifikationen“ bezeichnet (1960, S. 14); es zeigte sich, dass das leider zutraf.

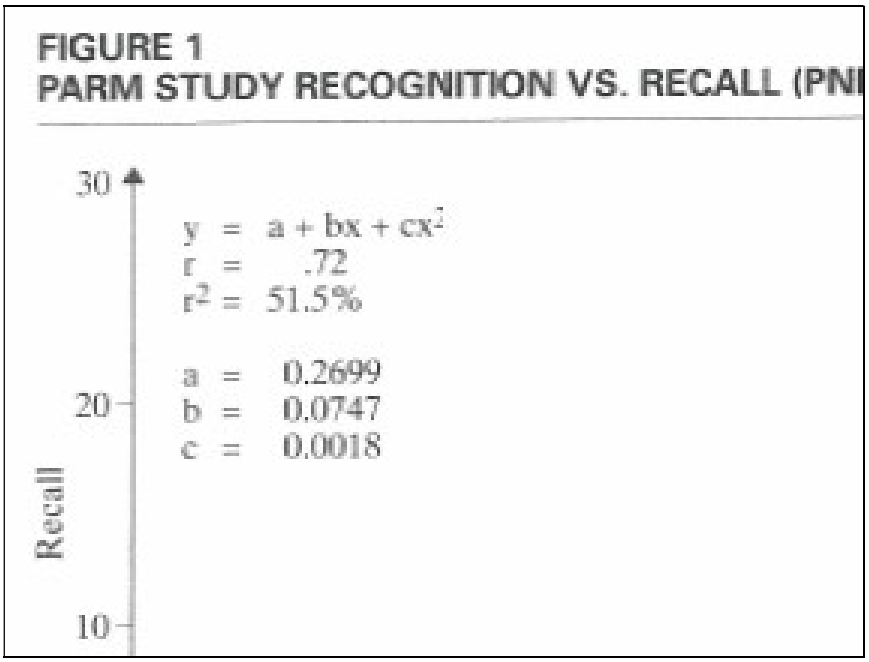
Es erwies sich nämlich als nicht möglich – auch nicht mit der freundlichen Unterstützung von ARF-Mitarbeiterinnen und der befreundeten ehemaligen Kollegin Anne Schuetz Zanes – die „Interessenten“- Ermittlung in der PARM-Untersuchung zu rekonstruieren, auf die sich Lucas bezogen hatte. Es liegen Auszählungen für Interessenten vor, aber nur auf wenige Produktfelder bezogen. Es war unmöglich, sie generell zu verwenden.

Selbstverständlich wies der PARM-Bericht Aufgliederungen für Männer und Frauen aus. Und diese lieferten erwartungsgemäss z.T. erheblich unterschiedliche Recognition-, Recall- und Interessen-Ergebnisse. Bei den meisten Anzeigen war mit ziemlicher Sicherheit zu erraten, ob sie sich primär an Männer oder Frauen wandten, oder an beide. Deshalb schien es möglich zu sein, das Geschlecht als „Interessenten“ - Ersatzindikator zu verwenden.

Dementsprechend wurde ein transatlantischer Versuch unternommen, alle 125 Testanzeigen in drei Gruppen aufzuteilen: Für Frauen, für Männer, für beide (oder unklar). Anschliessend führten wir separate Regressionsanalysen für Frauen und Männer durch. Aber diese erbrachten leider nicht Ergebnisse und Erkenntnisse, die über jene von Dodge hinausgingen. Ein arg einschränkender Faktor war hierbei die Raum-Zeit-Distanz von fast 5.000 km und 40 Jahren. So gab ich diesen Versuch auf.

Da Dodge seine Regressionsgrafiken nicht publiziert hatte, darf ich zwei von meinen als Ersatz anbieten.

Grafik A zeigt die Beziehung zwischen Recognition und Recall, bezogen auf alle Befragten. Da wir mehrere Formeln benutzten, können wir ausweisen, dass die optimale Beziehung nichtlinear ist, mit einer etwas höheren Korrelation als Dodge & Greene gefunden hatten (r =.72, erkl. Varianz 51,5%).

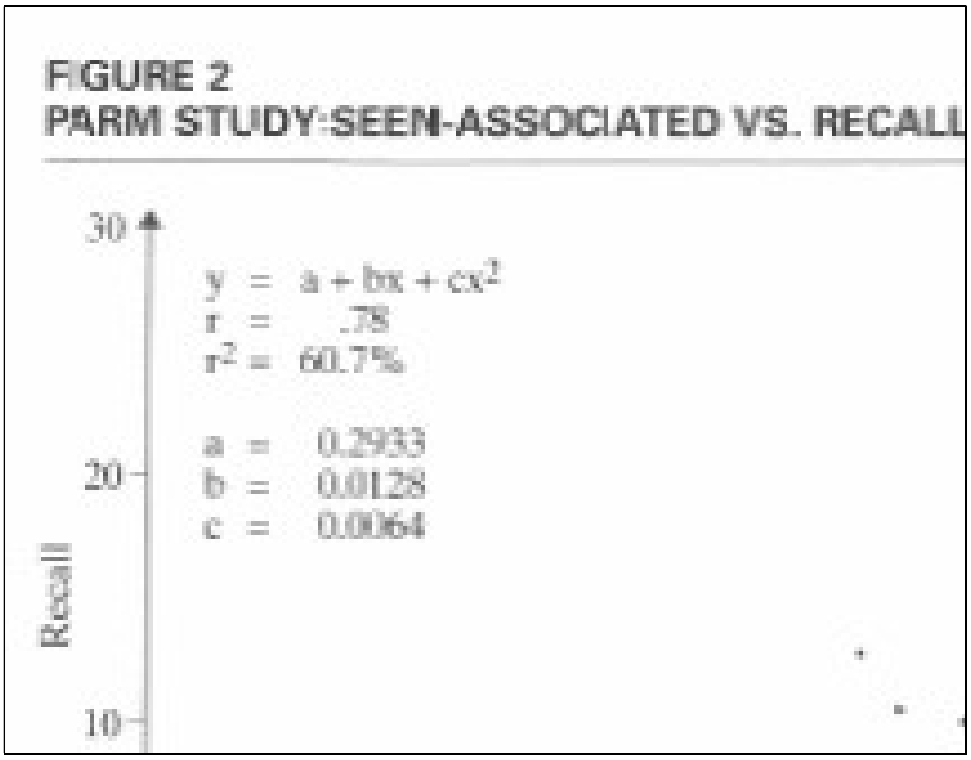


Jedoch ist diese Art von Beziehung als problematisch anzusehen. Recall, wie ursprünglich von Gallup & Robinson definiert, war nicht einfach irgendeine Art von Recall. Sondern die Befragten mussten „alles angeben, was sie von der genannten Anzeige erinnern konnten“; hierfür wurden verschiedene Fragen gestellt. Sofern die Befragten „die Anzeige korrekt identifiziert hatten, den Beweis geliefert hatten, dass sie ihren Inhalt erinnerten, bestätigt hatten, dass sie genau diese Anzeige meinten, und nicht in irgendeiner anderen Zeitschriften gesehen hatten“, dann wurden sie einbezogen unter der Messwert-Bezeichnung „Bewiesene Registrierung des Namens“ (”proved name registration”/ PNR). (ARF 1956, Vorwort, S.3)

Dieser verfeinerte „PNR“-(Recall-) Messwert kann nicht mit dem globalen Recognition-Wert in Beziehung gebracht werden, den Dodge verwendet hatte: Sofern ein(e) Leser(in) der »LIFE« - Ausgabe nur irgendeinen Teil der betr. Anzeige (z.B. das Bild), aber nicht den Namen der Marke bzw. des Werbungtreibenden beachtet hatte, wie könnte sie/ er sich bei der verfeinerten Recall - Ermittlung qualifizieren ?

Stattdessen muss man den Recognition-Messwert für die Marken- bzw. Firmennamen nehmen, der von Starch als ”seen - associated” eingeführt worden war: „Der Prozentsatz jener Leser der Ausgabe, die die Anzeige beachtet und dann erinnert haben, dass sie einen Teil der Anzeige gelesen hatten, in dem die Marke oder der Werbungtreibende identifiziert wurde.“ (ibid.)

Grafik B zeigt die Beziehung zwischen ”seen - associated” und „PNR“. Sie ist deutlich nichtlinear und weist eine höhere Korrelation auf (r =.78; erkl. Var. 60,7%).



Dementsprechend sollte in zukünftigen Analysen diese Beziehung zugrunde gelegt werden.

Zusammenfassung und Ausblick

Abschliessend ist demnach zu konstatieren, dass die in den 60er Jahren nach der PARM-Untersuchung der ARF geäusserte Kritik an der Recognition-Methode keine solide Grundlage hatte. Das bedeutet, dass die Verwendung der Recognition-Methode damals wie heute gerechtfertigt ist.

Offensichtlich hatten zwei der „Gründungsväter der Werbeforschung“, wie sie von Bartos und Pearson (1977) genannt wurden – nämlich Daniel Starch (als Urvater) und George Gallup – Forschungsmethoden entwickelt, die der praktischen Überprüfung durch die Anwender über Jahrzehnte hinweg, und nunmehr auch der methodologischen Prüfung ehrenvoll standgehalten haben. Überdies ist festzustellen, dass Sherwood Dodge, seinerzeit Vorsitzender des PARM-Komitees, mit seiner Analyse und Interpretation der Ergebnisse auf dem richtigen Weg war.

Die positiven Schlussfolgerungen dürfen nicht dahingehend missverstanden werden, Recognition-Messungen seien ohne Probleme. Mit diesen haben wir uns in diesem Bericht nicht befasst. Das Hauptproblem ist, dass Recognition von einer Anzahl unterschiedlicher Faktoren beeinflusst werden kann, von denen einige die Ergebnisse verfälschen können. Dieses Schicksal teilt Recognition mit Recall (s. Dodge-Ergebnisse)

und anderen Messverfahren, wie bekannt. Dieses Thema würde jedoch eine eigene, gründliche Auseinandersetzung erfordern.

Generell ist zu sagen, dass bei der Konzeption einer Recognition-Untersuchung bestimmte Regeln zu beachten und einzuhalten sind, die bereits von Starch und Gallup formuliert worden waren. Wie beide ausgeführt haben, kann jede Verkürzung oder Vereinfachung die Ergebnisse beeinflussen, sogar verfälschen.

Zur Zukunft ist aus meiner Sicht zu sagen, dass es vier Stufen für den Fortschritt der Forschung in Bezug auf Recognition und Recall gibt.

1. Die erste Stufe ist: Indem die Ergebnisse dieser beiden Messwerte (und anderer) in der richtigen Weise zueinander in Beziehung gesetzt werden, wie es Dodge vorgemacht hat, wird man weitere Informationen und Erkenntnisse gewinnen: Zu den Messwerten selbst wie auch zu den getesteten Anzeigen (oder anderen Werbemitteln).

2. Zweitens hat sich erneut die Bedeutung der relevanten Zielgruppen herausgestellt. In der zukünftigen Forschung muss jede Anstrengung unternommen werden, jene Personen zu ermitteln, die zu der/den speziellen Zielgruppe(n) gehören, um dann die Auswertung auf diese statt auf „alle Befragten“ zu beziehen.

3. Drittens hat sich wieder einmal gezeigt, dass der Faktor „Interesse/ subjektive Wichtigkeit/ persönliche Relevanz“ einer der wichtigsten Einflussfaktoren für Recognition und Recall ist.

   Dieser Befund erweitert unser Wissen über die Wirkungsweise der Werbung; er sollte bei zukünftigen Untersuchungen zur Werbewirksamkeit einbezogen werden.

4. Viertens wurden Informationen aus sehr verschiedenartigen Quellen zusammengetragen, und zu einem komplexen, trotzdem noch sehr unvollständigen Puzzle zusammengesetzt; aber es scheint sich bereits ein Bild abzuzeichnen. Auf diesem Wege ist weiterzugehen.

Post Scriptum: Was im Gehirn vor sich geht[2]

Nach Veröffentlichung des Aufsatzes kam eine amerikanische Untersuchung zur Verarbeitung von „echten“ und „falschen“ Gedächtnisvorlagen im Gehirn heraus (Schacter et al. 1966 / s. Degen 1997), die unsere Erkenntnis um eine wesentliche Stufe erweitert.

Bei der Untersuchung der Forscher an der Universität Harvard wurde 12 Versuchsteilnehmern 24 Listen mit inhaltlich verwandten Begriffen vorgelesen, und zwar in zwei Wellen. In der zweiten Welle wurde ein Teil der in der ersten Welle genannten Worte

[2] ergänzt 2001

benutzt, z. T. neue, also „falsche“. Es gab offenbar keine Warnung, dass vorher nicht benutzte Worte untergemischt worden waren. (Die Untersuchungsdetails sind im Original so kompliziert und für mich nicht ganz verständlich beschrieben, dass ich sie hier nicht zitiere. Für Interessenten ist die Quelle genau bezeichnet.)

Mittels Positronen-Emissions-Tomographie (PET) wurde während der 2. Welle gemessen, welche Gehirnregionen aktiviert wurden.

Ich zitiere nach Degen die Ergebnisse. Die Versuchsteilnehmer wiedererkannten 68% der „richtigen“ und 58% der „falschen“ Begriffe. Das ist nach dem, was wir gerade gelesen haben, nicht überraschend. Überraschend ist, dass die richtigen und die falschen Vorgaben im Gehirn unterschiedlich bearbeitet wurden.

Sowohl die richtigen wie die falschen Begriffe aktivierten den mittleren Schläfenlappen in der linken Gehirnhälfte. Dieser Abschnitt deckt sich weitgehend mit dem sogenannten Hippokampus, der immer in Aktion tritt, wenn Menschen auf frische Gedächtnisspuren zurückgreifen.

Die echten Vorgaben, nicht jedoch die falschen, aktivierten die Strukturen am Übergang zwischen Schläfen und Scheitellappen auf der linken Seite der Grosshirnrinde. Dort befindet sich ein sensorisches Sprachzentrum, in dem die lautlichen Einheiten (Phoneme) der gesprochenen Sprache entschlüsselt werden.

Die falschen Vorgaben aktivieren den präfrontalen Kortex im Stirnlappen und das Kleinhirn an der Rückseite des Schädels. Der präfrontale Kortex übernimmt im Gehirn sozusagen die Geschäftsführung, und zwar besonders dann, wenn Aufgaben nicht ohne weiteres bewältigt werden können. Am überraschendsten war die Aktivierung im Kleinhirn, dem bislang die Steuerung von Körperhaltung und Bewegungen zugesprochen wurde. Es mehrten sich aber Hinweise, dass es auch mentale Aufgaben erfüllt, z.B. Ausgleich von Defizite bei der Bewegungssteuerung.

Was bedeutet das für meine Reanalyse? Das Gehirn ist offenbar in der Lage, recht gut zwischen „richtigen“ und „falschen“ Vorgaben zu unterscheiden. Aber ob sich das bei einer bewussten Antwort in einem Interview als Identifikation von „richtig“ oder „falsch“ niederschlägt, liegt an der Fragestellung inkl. (k)einer Instruktion. Das zeigten die oben zitierten Tests von Neu bezüglich Zeitschriften-Anzeigen und in den anderen Tests mit Zeugen-Aussagen.

Irgendwann habe ich mal ein angebliches chinesisches Sprichwort betr. „Wissen“ gehört: Hinter den Bergen liegen immer noch Berge. Das trifft auch auf diesen neuen Forschungsansatz zu, denn nun stellt sich die Frage, wo und wie der Prozess der unterschiedlichen Behandlung von richtigen und falschen Vorgaben im Gehirn gesteuert wird und abläuft.

Danksagung
Der Autor dankt Betsy Proudfit und Roslyn Arnstein (ARF) für die Lieferung von PARM-Daten, Anne Schuetz Zanes für die Zusammenarbeit bei der PARM-Analyse; Jerome D. Greene und Prof. Darrell B. Lucas für wichtige Informationen vor der Abfassung dieses Aufsatzes; Dr. Valentine Appel, Dr. Herbert E.

Krugman, Dr. Abhilasha Mehta und Dr. William D. Wells für ihre teils kritischen, teils ermunternden Kommentare zu Entwürfen der englischen Version. Für kritische und hilfreiche Anmerkungen zur deutschen Fassung danke ich: Dr. Eva-Maria Hess, Dipl. Psych. Otfried Gerloff, Dr. Karlfritz Koeppler, Dr. Klaus Peter Landgrebe; letzterer war überdies sehr hilfreich, vom ursprünglich in englisch formulierten Bericht eine desanglisierte deutsche Version herzustellen.

Alle diese Hilfestellungen haben sehr wesentlich dazu beigetragen, dass diese Reanalyse deutlich verbessert wurde. Gewisse Unterschiede in den Auffassungen blieben selbstverständlich bestehen. Und der Autor trägt die volle Verantwortung für jegliche Irrtümer und Auslassungen.

ESOMAR wird sehr für die Zustimmung zur Veröffentlichung einer deutschen Fassung der von ihr veröffentlichten englischen Fassung gedankt. (In »marketing and research today«, May 1995)