Wenn Mediendarbietungen Antwortleistung und Antwortdisposition gleichzeitig bestimmen

Ursache-Wirkungs-Konvergenz als Problem der medien- und werbepsychologischen Forschung und Item-Response-Analyse als ein Lösungsvorschlag

Jens Woelke[1] & Steffen Kolb[2]

[1] Universität Leipzig, [2] HTW Berlin

ZUSAMMENFASSUNG

In der Medien- und werbepsychologischen Forschung übliche Analysemodelle untersuchen Effekte durch direkten Vergleich der Häufigkeiten oder Durchschnitte manifester Indikatorvariablen. Zentral für sogenannte True-Score-Analysen nach der klassischen Testtheorie (KTT) ist die Annahme, dass Schwankungen in den individuellen Ausprägungen der manifesten Indikatorvariable, die nicht auf Variationen der angenommenen Verursachungsgröße (wahrer ‚Prädiktor') zurückgehen sondern durch Drittvariablen (‚Fehler') verursacht werden, stochastisch unabhängig davon sind. Da diese Voraussetzung in der empirischen Praxis regelmäßig verletzt ist, kommen Kovarianz-, Interaktionseffekt- und Mehrebenenanalysen immer häufiger zum Einsatz. Die folgend diskutierte Re-Analyse von zwei medienpsychologischen Studien verweist theoretisch und empirisch auf eine Problematik, die in der Arbeit mit Häufigkeiten und Durchschnitten manifester Indikatorvariablen selbst bei Anwendung der oben genannten komplexen Analyseverfahren auftritt. Wertet man z. B. Wiedererkennungsaufgaben nach der Logik probabilistischer Testtheorien aus und trennt unter Anwendung der Signalentdeckungstheorie (SDT) zwischen der Antwortleistung sowie der Antwortdisposition, wird deutlich: die Drittvariable Antwortdisposition, verstanden als individuelle Neigung, in der Klassifikation von Objekten und Personen bestimmte Fehler zuzulassen und andere vermeiden zu wollen, ist mit der Antwortleistung verknüpft. Indem beide Aspekte situativ von Bedingungen abhängen, die das zu testende Medienangebot vorgibt, lässt sich der manifeste Indikator ‚Wiedererkennungsraten' nicht durch Kovarianzanalysen oder Forschungsdesigns wie Testwiederholungen bzw. Vergleich von Gruppen um die Einflüsse der Drittvariable ‚Antwortdisposition' bereinigen und als True-Score-Effekt interpretieren. Wenn Informationsangebote aus der Gesellschafts- und Wirtschaftskommunikation zunächst bestimmte Antwortdispositionen aktivieren und diese sodann das Ergebnis der Verarbeitung dieser Informationsangebote mitbestimmen, wenn also Ursache und Wirkung in introspektiven Daten konvergieren, sind Verfahren und Auswertungsmodelle wie die Signalentdeckungstheorie oder IRT-Modelle für die Analyse von eigentlich interessierenden Medieneffekten hilfreich.

Schlüsselbegriffe: Differentielle Werbewirkungsforschung, Item-Response-Analyse, Signalentdeckungstheorie, Diskriminationsleistung und Antwortneigungen

1 Ausgangspunkt: Testen und Messen nach klassischer versus probabilistischer Testtheorie

Informationsangebote öffentlicher Kommunikationsmedien ‚wirken' grundsätzlich nicht auf alle NutzerInnen gleich, sondern abhängig von Persönlichkeitseigenschaften sowie vom situativen Kontext des Medienhandelns. Diese Annahme ist Konsens in der medienpsychologischen wie in der Werbeforschung (vgl. Suedfeld & Tetlock, 2003; Richter 2007; Woelke, 2008) – ungeachtet der Tatsache, dass die Frage nach durchschnittlichen kausalen (Werbe)Effekten[1] aus medienethischen, medienpolitischen oder medienökonomische Gründen nach wie vor relevant ist und bearbeitet wird (vgl. Russel, 2002; Mackay, Ewing, Newton & Windisch, 2009; Dardis, Schmierbach & Limperos, 2012). Zwei Gründe lassen die Kritik an medien- und werbepsychologischen Projekten mit Begrenzung auf nur durchschnittliche kausale (Medien)Effekte jedoch gerechtfertigt erscheinen: Medienwirkungen entstehen nicht allein durch additive Verknüpfung von zwei oder mehr Einflussfaktoren, weshalb sich Einflüsse einer einzelnen, als Verursachungsbedingung angenommenen Größe ohne simultane Beobachtung anderer erheblicher Einflussgrößen (‚Drittvariablen') kaum valide beurteilen lassen und Kovarianz- und Moderatoranalysen der Auswertung getrennter Gruppen vorzuziehen sind (vgl. Cohen, Cohen, West & Aiken, 2002).

[1] Zur Unterscheidung von durchschnittlichen und individuellen kausalen Effekten siehe Neyman (1990).

Allerdings sind die ‚anderen Einflussfaktoren' in Kovarianz- oder Moderatoranalysen (vgl. Nicovich, 2005; Schemer, 2007; Klein, 2009; Slater, Hayes, Reineke, Long & Bettinghaus, 2009; Taylor, Strutton & Thompson, 2012) meist nur quasi-experimentelle Faktoren, sodass Verteilungsvoraussetzungen und die Annahme der Unkorreliertheit von ‚True Score' und Fehlern in der klassischen Testtheorie nicht nur in Beobachtungsstudien (Steyer, Partchev, Kroehne, Nagengast & Fiege, 2010), sondern regelmäßig auch in experimentellen Analysen nicht erfüllt sind.

Abgesehen davon, dass die Berücksichtigung von Interaktionseffekten und von Verteilungsvoraussetzungen quasi-experimenteller Faktoren einen methodischen Fortschritt bedeutete, findet ein zumindest bekanntes und relevantes Problem empirischer Forschung damit noch immer eher geringe Beachtung – es wird folgend unter dem Stichwort ‚Ursache-Wirkungs-Konvergenz' adressiert. Angesprochen ist der in Item-Response-Theorien (IRT) bzw. der Latent-Trait-Analyse (LTA) bearbeitete Umstand, dass ein Testwert Y (z. B. ‚Wiedererkennungsraten') nicht nur den um eine Fehlerkomponente ergänzten ‚True Score'-Effekt X (z. B. den tatsächlichen Gedächtniseffekt eines Informationsangebotes) repräsentiert, sondern neben der vom Informationsangebot bestimmten (zunächst latenten) Fähigkeit X1 einer Person (z. B. zuvor dargebotene Reize wiedererkennen zu können) zugleich auch deren Antwortdisposition X2 misst (Steyer & Eid, 2001).

Diese Annahme bedeutet einen wesentlichen Unterschied zu Analysen nach der klassischen Testtheorie (KTT): Antwortdispositionen wie z. B. das Verwechseln von Namen, auch wenn man sich Personen grundsätzlich gut merken kann, die Neigung auf schwierige, Unsicherheit erzeugende Aufgaben eher mit ‚raten' oder 'keine Antwort geben' zu reagieren, oder bestimmte Fragen eher zu verneinen als zu bejahen, werden in Haupteffektanalysen und mit der Wiedererkennungsrate als manifester Indikator über die Annahme einer Unkorreliertheit von Fehlern und ‚True Score' durch Vergleiche zwischen Gruppen (unter der Voraussetzung homogener Fehlervarianzen) oder mit wiederholten Beobachtungen derselben Person quasi ausgeblendet oder in der Moderatoranalyse in einen separaten und einen gemeinsamen Erklärungsanteil zerlegt (wobei es im Fall des Recognition-Tests keine essentielle Multikollinearität zwischen Wiedererkennungsraten und Antwortneigung geben dürfte).

In der IRT/LTA liegt der Fall etwas anders: Bei der Ableitung empirisch testbarer Bedingungen kommt man auch hier nicht ohne Unabhängigkeitsannahme aus, allerdings wird keine generelle, sondern nur eine bedingte stochastische Unabhängigkeit angenommen. Bedingte oder lokale stochastische Unabhängigkeit bedeutet, dass die Informationsfunktion eines Tests maximal ist, wenn dessen Schwierigkeit bzw. die damit wahrscheinliche individuelle Antwortdisposition mit der latenten Fähigkeit oder anderen latenten Eigenschaften einer Person (θ) übereinstimmt bzw. abnimmt mit größer werdender Differenz zwischen Schwierigkeit der Testaufgabe und latenten Lösungsfähigkeit. Am Beispiel einer Skala wie der CSII-D (vgl. Woelke & Dürager, 2012) aus der Persuasionsforschung, die eine Anfälligkeit für interpersonale Beeinflussungsversuche misst und sich in der Prognose differentieller Werbeeffekte als zielführend erwiesen hat (vgl. Woelke, 2008), bedeutet das: Die individuellen Messwerte auf den einzelnen Items der CSII-D geben nicht generell, sondern abhängig von den Ausprägungen der latenten Eigenschaft ‚Beeinflussbarkeit' (X1) einer Person auch die Antwortdispositionen X2 von Befragten wieder, diese Items z. B. deutlicher ‚abzulehnen'. Befragte mit einer etwas höheren latenten Ausprägung von Beeinflussbarkeit (in Tabelle 1 als θ bezeichnet) sehen die Anfälligkeit für interpersonale Beeinflussung u.U. gar nicht als ‚Problem' und stimmen CSII-Items ‚eher zu'. Personen mit einer etwas geringeren latenten Ausprägung von Beeinflussbarkeit fassen die Anfälligkeit für interpersonale Beeinflussung dagegen möglicherweise als individuelle Eigenschaft auf, die sozial negativ gesehen wird – entsprechend dürften diese Personen CSII-Items deutlicher ablehnen, als aufgrund ihrer latenten Ausprägung von Beeinflussbarkeit vorhersagbar. Wenn Items eines Tests in der sozialen Realität über die Bandbreite von manifesten Antworten tatsächlich nicht gleich gut messen, ist es zielführend, die Informationsfunktion von Items zu ermitteln.

Tabelle 1: Item-Informations-Werte der CSII-D (Woelke & Dürager 2012) im Graded-Response Model.

Item Information Function Values for Group 1 at 15 Values of θ from -2.8 to 2.8 (Back to TOC)

Item	Label	θ: -2.8	-2.4	-2.0	-1.6	-1.2	-0.8	-0.4	-0.0	0.4	0.8	1.2	1.6	2.0	2.4	2.8
1	beeinfluss01	0.08	0.11	0.14	0.18	0.23	0.26	0.29	0.30	0.31	0.31	0.31	0.31	0.31	0.31	0.31
2	beeinfluss02	0.01	0.03	0.07	0.17	0.37	0.73	1.13	1.31	1.27	1.32	1.40	1.40	1.41	1.35	1.31
3	beeinfluss03	0.02	0.04	0.10	0.21	0.44	0.79	1.13	1.26	1.27	1.31	1.30	1.30	1.29	1.23	1.24
4	beeinfluss04	0.01	0.02	0.04	0.09	0.18	0.36	0.63	0.92	1.07	1.07	1.09	1.11	1.08	1.10	1.17
5	beeinfluss05	0.04	0.07	0.14	0.26	0.45	0.68	0.87	0.94	0.97	0.98	0.97	0.97	0.97	0.94	0.87
6	beeinfluss06	0.01	0.02	0.04	0.09	0.18	0.36	0.62	0.90	1.06	1.10	1.13	1.14	1.13	1.08	1.02
7	beeinfluss07	0.00	0.01	0.03	0.07	0.15	0.34	0.69	1.12	1.38	1.42	1.46	1.48	1.47	1.41	1.37
8	beeinfluss08	0.02	0.04	0.07	0.13	0.23	0.38	0.56	0.72	0.80	0.83	0.83	0.83	0.81	0.77	0.73
9	beeinfluss09	0.10	0.16	0.25	0.36	0.47	0.55	0.59	0.60	0.60	0.60	0.59	0.58	0.57	0.55	0.55
10	beeinfluss10	0.17	0.17	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.15	0.15	0.15	0.14	0.13	0.12
11	beeinfluss11	0.25	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.25	0.25	0.25	0.24	0.23	0.20	0.17
12	beeinfluss12	0.29	0.29	0.30	0.30	0.30	0.29	0.29	0.29	0.29	0.29	0.29	0.28	0.27	0.25	0.21
	Test Information:	1.99	2.21	2.60	3.28	4.43	6.17	8.21	9.78	10.43	10.64	10.79	10.81	10.66	10.31	10.07
	Expected s.e.:	0.71	0.67	0.62	0.55	0.47	0.40	0.35	0.32	0.31	0.31	0.30	0.30	0.31	0.31	0.32

Marginal Reliability for Response Pattern Scores: 0.86

Nur so kann beurteilt werden, wie gut oder schlecht manifeste Indikatorvariablen (z. B. ‚Wiedererkennungsraten') das interessierende Phänomen (das fragliche Gedächtnispotential eines Informationsangebotes) bei bestimmen Ausprägungen der zunächst latenten Fähigkeit einer Person (z. B. Reize potentiell wiederzuerkennen) messen – Auswertungen gemäß IRT-Modellen wie dem von Rasch, Mokken oder anderen (vgl. Mokken, 1997; Scheiblechner, 2007) liefern die dazu nötigen Informationen.

2 Warum man bei Werbewirkungstests zwischen Aufgabenschwierigkeit und Personenfähigkeit trennen sollte – zur situativen Abhängigkeit von Antwortleistung und Antwortdisposition in verbalen Daten

Für die Anwendung des Konzepts lokaler stochastischer Unabhängigkeit gibt es mehrere Gründe: In der Bildungsforschung werden probabilistische Testmodelle (z. B. Rasch-logistisches Modell, 2PL oder GRM) nicht nur aus forschungsökonomischen Gründen angewendet sondern vor allem aufgrund der Tatsache, dass sich von den vielen Aufgaben, die valide Schulleistungstest erfordern, tatsächlich nur eine begrenzte Anzahl praktisch umsetzen lässt. Zur Auswahl geeigneter Testaufgaben ist allerdings die Information unerlässlich, wie schwierig deren Lösung in welchen Bereichen latenter Personenfähigkeit ist. Gemäß dieser Information können Aufgaben ausgewählt werden, die in einem definierten Fähigkeitsbereich hinreichend schwierig sind und dort feindifferenziert messen (zielführend z. B. bei Auswahltests unter Piloten) oder Aufgaben, die jeweils nur mittelmäßig schwierig sind, dafür aber einen breiten Bereich von Personenfähigkeit abdecken (notwendig etwa in der Berufsberatung) (vgl. Kubinger 2005).

Man könnte einwenden, dass die methodisch aufwändige und voraussetzungsreiche Trennung von Aufgabeschwierigkeit und (latenter) Fähigkeit einer Person eine Fingerübung darstellt, die im Grunde nur dann relevant wird, wenn Tests erhebliche Konsequenzen für die Handlungsfreiheit von Menschen haben, z. B. in der angesprochenen Schulleistungsforschung. So gesehen wäre der praktische Nutzen des Konzepts lokaler stochastischer Unabhängigkeit für die medien- und werbepsychologische Forschung fraglich, wo solche Entscheidungen nicht getroffen werden und mit dem Ansatz der Moderatoranalyse mittlerweile recht differenzierte Erkenntnisse möglich sind.

Studien aus anderen Forschungsbereichen zeigen aber, dass die Trennung von (latenter) Personenfähigkeit und (spezifischer) Aufgabenschwierigkeit höchst bedeutsam ist, wenn diese zwei Aspekte eine gemeinsame, im Test selbst liegende Ursache haben. In der Intelligenzforschung z. B. wird aktuell diskutiert, in wie weit mentale Geschwindigkeit und mentale Kapazität als zwei korrelierende Aspekte (Sheppard & Vernon, 2007) unabhängig sind und – gemäß gängiger Intelligenztestmodelle (z. B. BIS und APM; vgl. Neubauer & Knorr, 1998) – tatsächlich separat oder extern gemessen werden können. Anhand mehrerer fMRT-Studien kommen Rypma und Prabhakaran (2009) zu dem Schluss, dass weder ‚Kapazität' noch ‚Geschwindigkeit' ‚Trait'-Eigenschaften repräsentieren, sondern beide im Grunde ‚State'-Merkmale[2] sind, die vom jeweils anderen, als Mediator fungierenden Merkmal abhängen. Mit anderen Worten: Mentale ‚Kapazität' ist ein latenter Teilaspekt von Intelligenz, dessen Beitrag zu einem manifesten IQ-Testwert von den Bedingungen abhängt, die der Test in Bezug auf mögliche Ausprägungen des anderen (latenten) Teilaspekts mentale ‚Geschwindigkeit' vorgibt.

Folglich halten Partchev und de Boeck (2012) die sogenannte ‚cognitive correlates method' als wenig geeignet zur Kontrolle von IQ-Testwerten und schlagen aus der IRT abgeleitete Auswertungsmodelle vor. Ähnliche Hinweise finden sich in der Persönlichkeitsforschung: In der Diskussion um Aggressionsreaktionen nimmt man aktuell an, dass deren Messung über manifeste Indikatoren zugleich eine individuelle Traiteigenschaft ‚Agressivität' (im Sinne von Fähigkeit') und als auch das Statemerkmal ‚situationale Sensitivität' (als personenspezifische Interpretation einer Situation) erfasst (Schmitt et al., 2008).

Die Möglichkeit, einen Teilaspekt von Handlungen (z. B. komplexe Aufgaben schnell versus korrekt versus vollumfänglich zu lösen) als separates Merkmal unkonfundiert zu erfassen und den Einfluss auf den jeweils anderen Teilaspekt per Moderatoranalyse oder Kovarianzanalyse zu kontrollieren, fehlt aber nicht nur bei Intelligenz- oder Persönlichkeitstests: Im ‚Eurobarometer' finden sich regelmäßig Länderunterschiede, z. B. bezüglich des Interesses an aktuellen wissenschaftlichen Entdeckungen und technologischen Entwicklungen[3]. Studien wie diese sind ein typischer Fall, in dem die Antworten von Befragten nicht nur ihre (latenten) Einstellungen bezüglich des fraglichen Sachverhaltes wiedergeben, sondern zugleich auch generelle Antwortdispositionsunterschiede abbilden (z. B. dass Befragte aus Österreich regelmäßig kritischer sind, d.h. seltener Zustimmung äußern als Befragte aus der Schweiz oder Deutschland). Im Ländervergleich beobachtete Unterschiede hinsichtlich manifester Indikatoren sind daher noch vorsichtiger zu interpretieren als ohnehin angeraten wird – eine Forderung, die über die gängigen Analysen von Konstrukt- und Itemäquivalenz hinausgeht (Kolb & Beck, 2011).

Das im vorliegenden Beitrag als Ursache-Wirkungs-Konvergenz bezeichnete Problem, das beide für das Zustandekommen eines manifesten Testwertes ursächliche Aspekte, also nicht nur die (latente) Fähigkeit einer Person, ‚korrekt' antworten zu können (hier als Antwortleistung adressiert), sondern auch die Schwierigkeit einer Testaufgabe (hier Antwortdisposition genannt) State-Merkmale sein könnten, sollte auch in der medien-, markt- und werbepsychologischen Forschung nicht unterschätzt werden.

[2] Zur Unterscheidung von Personeneigenschaften in ‚Trait' oder ‚State' in der medienpsychologischen Forschung siehe Schmitt (2004) und s.u.

[3] Ergebnisse im Eurobarometer 2010 (vgl. OQ1, 2010): Anteil der sehr interessierten Personen: Österreich = 21 Prozent, Deutschland = 32 Prozent, Schweiz = 33 Prozent oder für den Informationsstand über Wissenschaft und Technik (Anteil der Personen die angeben, schlecht informiert zu sein: Österreich = 51 Prozent; Schweiz = 35 Prozent; Deutschland = 36 Prozent).

Ob die übliche Annahme der persönlichkeitspsychologischen Forschung (vgl. Schmitt, 2004), das zentrale Eigenschaften einer Person und damit auch deren Antwortdisposition(en) zeitlich und über verschiedene Handlungen hinweg stabile Merkmale (‚Trait') sind, deren Realisation als konkreter Zustand (‚State') im Grunde die Form einer Konstante aufweist, auch für die Antwortdisposition in Gedächtnistests gilt, wird in einer Re-Analyse von zwei Studien zur Werbewirksamkeit verschiedener Informationsdarbietungen geprüft.

3 Zur Unterscheidung von Antwortleistung und Antwortdisposition in Gedächtnistest

Gedächtnisleistungen gehören neben Einstellungen, Emotionen und Verhaltensdispositionen zu den zentralen Analyseebenen in der Werbewirkungsforschung. Nach einer Metaanalyse von Wirth und Kolb (vgl. 1999) über 255, zwischen 1970 und 1997 in 11 deutschsprachigen und internationalen kommunikationswissenschaftliche Fachzeitschriften erschienen Studien setzen sowohl Rezeptions- als auch Wirkungsforschung schwerpunktmäßig Gedächtnismessungen ein: 52 Prozent der erhobenen Indizes sind gestützte oder ungestützte Recall-Tests oder Recall-Test-Mischtypen, ein weiteres Zehntel der untersuchten Studien verwendeten unterschiedliche Retrieval-Typen. Auch die Aktualisierung der Studie ergibt ähnliche Ergebnisse, wenngleich die Werte zu Gunsten von Recognition-Tests zurückgehen: Im Zeitraum von 1998 bis 2005 erheben die Autoren nur noch 36 Prozent reiner Recall-Messungen und sieben Prozent Mischtypen (vgl. Wirth, Heydecker & Kolb, 2006). In der Praxis der Marktkommunikation ist ‚Gedächtnis' ebenfalls eine zentrale Ziel- und Planungsgröße (vgl. Engelhardt, 1999) und auch hier sind die Gedächtnismessungen meist eindimensional konzipiert: Die Testergebnisse werden unisono als Medienwirkung bzw. kommunikative Leistung des Informationsangebotes gewertet. Zwischen der Antwortleistung und den Bedingungen ihres Zustandekommens wird meist nicht unter-schieden – ein Umstand, der angesichts des Verbreitungsgrads von Gedächtnistests mit Erinnerungshilfen (gestützter Recall oder Recognition-Test) verwundert.

Die Anwendung der Signalentdeckungstheorie (SDT) auf Wiedererkennungsdaten macht die aus der Anwendung der KTT folgende Bedingung obsolet, dass andere für das Zustandekommen der individuellen Testwerte ebenfalls verantwortliche Drittvariablen (hier die ‚Antwortneigung' verstanden als individuelle Disposition, in der Entdeckung oder Klassifizierung von Objekten und Personen Fehler eher zuzulassen oder eher vermeiden zu wollen) zufällig verteilt und von der vermeintlichen Ursache (hier dem Gedächtnispotential eines Informationsangebotes) stochastisch unabhängig sein müssen.

Um die Signalentdeckungstheorie anwenden zu können, ist ein Wiedererkennungstest a) als ja/nein-Recognition zu konzipieren und sind b) die Antworten der Befragten zu den zwei Typen von Itemvorgaben (‚alte' Items = im medialen Angebot bereits vermittelte Reize und ‚neue' Items = erstmals im Wiedererkennungstest präsentiert) in einem Vier-Felder-Schema aufzuschlüsseln: Das Wiedererkennen eines im Medienangebot zuvor tatsächlich gezeigten Reizes ist ein ‚Treffer' und dessen Nicht-Wiedererkennung ein ‚Verpasser', das Wiedererkennen eines zuvor nicht gezeigten Reizes ein ‚falscher Alarm' und das richtige Nicht-Wiedererkennen desselben eine ‚korrekte Zurückweisung' (siehe Kategorien in Abbildung 1). Die SDT beschreibt Gedächtnisleistungen als Funktion der Empfindungsstärke eines zu erinnernden oder wiederzuerkennenden Items (vgl. Velden, 1982). Beim Vergleich mehrerer Personen oder alternativ von mehreren Reizen streuen die Empfindungsstärken gemäß dem Modell des sensorischen Kontinuums um eine mittlere Empfindungsstärke. Werden in einem Wiedererkennungstest nicht nur ein zuvor bereits präsentierter Reiz (‚altes Item'), sondern ein zusätzlicher, bisher nicht präsentierter Reiz vorgelegt (‚neues Item'), streuen die Empfindungsstärken in zwei separaten Verteilungen (mit μ 1 für ‚neue' und μ 2 für ‚alte' Items). Im Idealfall, z. B. wenn Reize eindeutig unterscheidbar sind oder bei Medienangeboten mit wenigen ausgewählten Informationen, sind die beiden Verteilungsfunktionen (nahezu) überlappungsfrei.

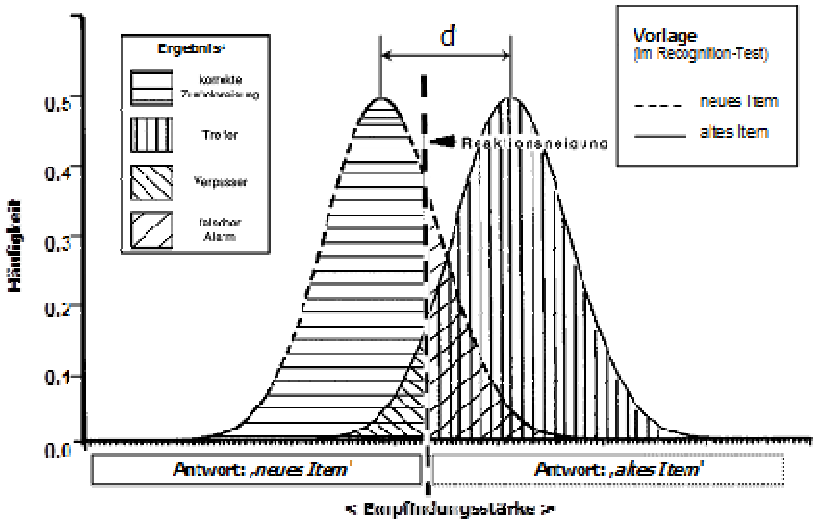


Abbildung 1: Ergebnisse im Wiedererkennungstest nach dem Modell des sensorischen Kontinuums.

Im Alltag jedoch, wenn Personen komplexe multimediale und multimodale Informationsangebote mit vielen ähnlichen Reizen wenig aufmerksam rezipieren oder die Abfrageliste eines Gedächtnistests viele Items enthält, überlagern sich die Verteilungsfunktionen für die Empfindungsstärken von ‚alten' und ‚neuen' Items. Löst nun ein ‚altes' Item, d.h. bereits zuvor dargebotener Reiz einen kleineren Empfindungswert aus als ein ‚neues' Item, kommt es zu fehlerhaften Zuordnungen: Der neue, zuvor nicht gezeigte Reiz wird fälschlich wiedererkannt (‚falscher Alarm'), der zuvor tatsächlich gezeigte Reiz jedoch übersehen (‚Verpasser'). Die Fähigkeit von Personen, zwischen ‚alten' und ‚neuen' Items unterscheiden zu können – sie ist z. B. abhängig von der Stärke einer Reizdarbietung und damit ein Maß für das Aufmerksamkeits- bzw. Gedächtnispotential eines Medienangebotes – wird in der SDT als Diskriminationsleistung bezeichnet.

Die Entscheidung, ein Item in der Vorlagenliste des Wiedererkennungstests als ‚alt' oder als ‚neu' zu markieren, ist aber nicht nur eine Frage der Diskriminationsfähigkeit. Die Empfindungsstärke, die Reize bei wenig aufmerksamer Rezeption, bei einer Vielzahl von Informationen im Medienangebot und/oder bei einer großen Anzahl von Items in der Abfrageliste des Wiedererkennungstest hervorrufen, lässt selten nur eine, sondern meist zwei in gleicher Weise plausible Entscheidungen zu - jene, dass der Reiz in der vorherigen Medien-darbietung bereits vorkam (‚altes' Item), aber auch die gegenteilige Entscheidung, dass es sich um einen bisher nicht präsentierten Reiz (‚neues' Item) handelt. Den Wert der Empfindungsstärke, ab dem sich Personen bei Unsicherheit für die Antwort ‚altes' Item statt ‚neues' Item entscheiden, markiert die in Abbildung 1 als Reaktionsneigung bezeichnete senkrechte Linie. Die Reaktionsneigung von Personen (IRT = Aufgabenschwierigkeit, hier auch als Antwortdisposition bezeichnet) lässt sich zunächst als individuell ‚festgelegt' vorstellen:

Aufgrund bestimmter Prädisposition (vermutlich handelt es sich um offene, spontane und selbstbewusste Menschen mit geringer externaler Kontrollüberzeugung) neigen manche Personen grundsätzlich dazu, Reize auch bei kleiner Empfindungsstärke als ‚alt' zu markieren, während andere Personen (vermutlich ängstlichere und weniger selbstbewusste Menschen mit höherer externaler Kontrollüberzeugung) solche Reize eher als ‚neu' einstufen. Vorstellbar ist aber auch, dass die Ursachen für unterschiedliche Reaktionsneigungen situativ sind: Wer in einem Quiz mit einer richtigen Antwort eine Runde weiter kommt, wird sich frühzeitig äußern, auch wenn subjektiv noch große Unsicherheit über die Korrektheit der Antwort besteht. Wenn eine falsche Antwort dagegen Punktabzug bedeutet, den Kontrahenten einen Punkt bringt oder in Spielshows wie MILLIONENSHOW oder WER WIRD MILLIONÄR sogar zum Verlust des Gewinns führt, warten Quizteilnehmer eher länger mit ihrer Antwort ab - und zwar so lange, bis sie sicherer sind, damit richtig zu liegen.

4 Informationsgehalt der SDT

Das ‚Wiedererkennen' von zuvor nicht präsentierten (‚falscher Alarm') bzw. das ‚Übersehen' von zuvor präsentierten Reizen (‚Verpasser') steht offenbar nicht nur im Zusammenhang mit der kognitiven Leistungsfähigkeit – wenn die Entscheidung für die Antwortkategorien ‚alter Reiz' versus ‚neuer Reiz' unter Unsicherheit fällt, geben Wiedererkennungsraten ebenso die Antwortdisposition von Personen wieder. Wie Abbildung 2 verdeutlicht, ist deren Ausprägung von erheblicher Konsequenz für das Ergebnis einfacher Gedächtnistests: Bei der Antwortdisposition ‚progressives Rating' ist die Wahrscheinlichkeit geringer, in einem Medienangebot tatsächlich dargebotene Reize im Wiederkennungstest als ‚nicht wiedererkannt' zu markieren (‚Verpasser'). Allerdings ist dadurch die Gefahr größer, dass zuvor nicht dargebotene Reize fälschlich wiedererkannt werden (‚falscher Alarm') – die Wiedererkennungen sind wenig korrekt. Umgekehrt stellt sich die Situation für Personen mit der Antwortdisposition ‚konservatives Rating' dar: Die Wahrscheinlichkeit, im Medienangebot nicht enthaltene Stimuli fälschlich als ‚wiedererkannt' zu markieren (‚falscher Alarm') ist wesentlich geringer. Allerdings ist bei einem ‚konservativem Rating' die Gefahr größer, im Medienangebot tatsächlich enthaltene Stimuli als ‚nicht wiedererkannt' einzuordnen, d.h. einen ‚Verpasser' zu erzielen.

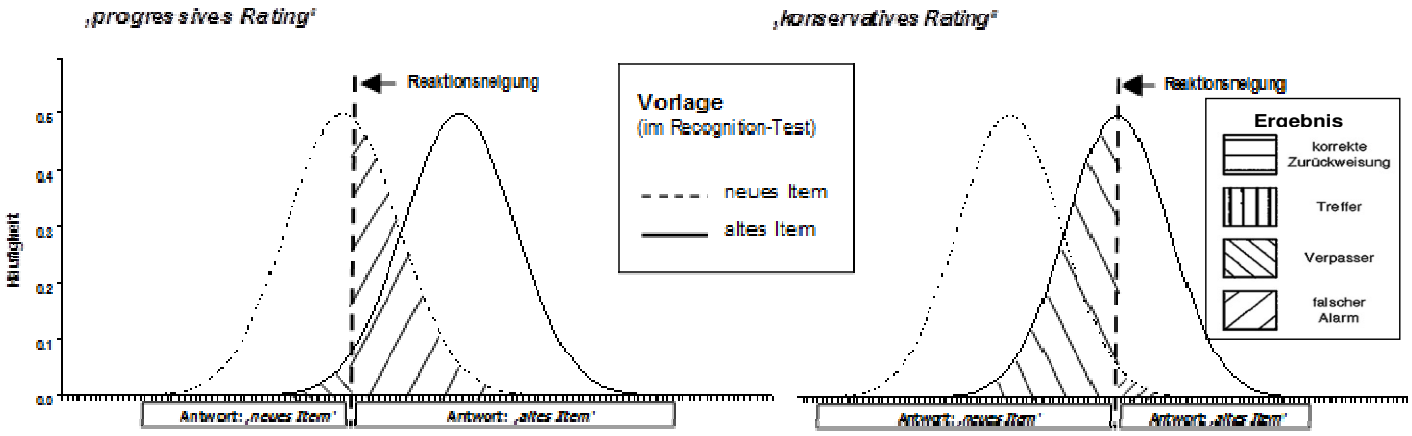


Abbildung 2: Reaktionsneigung und Fehler in der Wiedererkennung: ‚progressives versus konservatives Rating.

Mit der Unterscheidung von Diskriminationsleistung und Reaktionsneigung in der Signalentdeckungstheorie wird deutlich, vor welchem Problem die medien- und werbepsychologische Forschung steht, die allein Wiedererkennungsraten (Anzahl der als wiedererkannt markierten, im Medienangebot tatsächlich präsentierten Stimuli) bzw. die Anzahl von Nennungen in einem Recall-Test als Gedächtnisleistung begreift und darüber Aussagen über das kognitive Potential von Medienangeboten trifft. Unbestritten ist, dass hohe Wiedererkennungsraten bzw. viele Nennungen in einem Recall-Test zunächst auf ein erhebliches Aktivierungspotential von Medienangeboten hinweisen; ebenso ist es aber auch möglich, dass Befragte einen Gedächtnistest ‚progressiv' bearbeiten, d.h. Antworten geben, ohne sich über das tatsächliche Rezeptionserleben ‚sicher' zu sein. Umgekehrt lassen sich geringe Wiedererkennungsraten oder wenige Nennungen in einem Recall-Test als minimales Aktivierungspotential von Medienangeboten deuten; allerdings kann die Ursache auch hier im Antwortverhalten (= Antwortdisposition) liegen: Das Medienangebot kann genauso viel Aktivierungspotential entfaltet haben wie im Fall zuvor, nur haben die Befragten im Gedächtnistest vielleicht ‚konservativer' entschieden.

5 Antwortleistung und Antwortdisposition als ‚State'-Variablen in Wiedererkennungstests – eine Reanalyse von zwei medienpsychologischen Studien

Um die empirische Relevanz der theoretisch plausiblen Unterscheidung von Messwerten in einen Leistungsaspekt und eine Antwortdisposition für die medien- und werbepsychologische Forschung zu verdeutlichen, wurden zwei frühere Medienwirkungsstudien reanalysiert. In beiden Studien wurden Gedächtniseffekte über ja/nein-Wiedererkennungstests erfasst – entsprechend lassen sich unter Anwendung der SDT die tatsächliche Wiedererkennungsleistung (IRT = Personenfähigkeit) und die Antwortdisposition (IRT = Aufgabenschwierigkeit) einer Person beim Lösen der Wiedererkennungsaufgabe getrennt berechnen und ausweisen.

5.1 Skizze der Ursprungsstudien mit Wiedererkennungsaufgaben

Studie 1, eines von drei Experimenten in der Studienreihe zur Analyse kommunikativer Abgrenzungen von Zuschauern am Beispiel Programmintegration (vgl. Woelke, 2004), wurde im Wintersemester 2000 an der Universität Jena durchgeführt. Es handelte sich um ein Zwei-Gruppen-Experiment mit zufälliger Zuteilung der Untersuchungsteilnehmer[4]: In beiden Gruppen war Stimulus eine aus Originalszenen und Originalwerbespots erstellte Fassung der Show DIE DICKSTEN DINGER (RTL 2, 1997-2001), die von drei Werbeblöcken unterbrochen wurde und wie in der Originalversion Moderation, Spotpräsentationen, Charts und ein Gewinnspiel enthielt. In der ersten Experimentalgruppe (‚Sendungsgruppe') waren die sechs Zielwerbespots (GMX-Internet, IKEA-Möbel, CONTINENTAL-Reifen, TUI-Flugreisen, SWATCH-IRONY-Uhren und TUBORG-Bier) innerhalb der Sendung zu sehen, abgestimmt auf die Anmoderation.

Für die zweite Experimentalgruppe (‚Werbungsgruppe') wurden die Stimuli rotiert: Hier waren die sechs Zielspots Element der Werbeblöcke, welche die Sendungen zwischen den drei Sendungsteilen unterbrachen. Die Position der Zielspots, die in der Stimulusfassung für die Sendungsgruppe als Teil der Show DIE DICKSTEN DINGER nach der Anmoderation ausgestrahlt wurden, übernahmen jene Spots, die in der ersten Stimulusfassung die Stellen im Werbeblock ausfüllten, die hier in Stimulusfassung zwei von den sechs Zielspots eingenommen wurden.

Studie 2 war eine ebenfalls experimentelle Untersuchung zur Wirksamkeit von Product Placement, die im Frühjahr 1996 an der Freien Universität sowie an der Technischen Universität Berlin durchgeführt wurde (vgl. Woelke, 1998). Die Untersuchungsteilnehmer [5], per Zufallsverfahren aufgeteilt in zwei Gruppen, sahen einen Ausschnitt (Dauer ca. 55 Minuten) des Spielfilms BIS ANS ENDE DER WELT (Regie: Wim Wenders, 1987). In diesem Film waren zahlreiche Produkte und Marken zu sehen, die aktuelle Angebote aber auch Zukunftstechnologien bekannter Marken (wie BOSS, BAUKNECHT, COCA COLA, LUFTHANSA, SONY-CAR-INFORMATION-SYSTEM, SONY-MOBILE-BILD-TELEFONE; SHARP-NETBOOK) vorstellten: Gruppe A sah den Film in der Originalfassung mit Product Placements, allerdings mit geänderter Anzahl von Platzierungen: Einblendungen von Produkten und Marken aus der realen Konsumwelt (BOSS, BAUKNECHT und COCA-COLA) wurde entfernt, ebenso einige sehr kurze Einblendungen von LUFTHANSA, SONY (2x: Mobil-Bild-Telefon und Car-Informationssystem) und SHARP, die zusätzlich zu langen und ausreichend erkennbaren Darbietungen für diese vier Produkte/Marken in so genannten Product-Placement-Inseln im Spielfilm vorkamen. In der Stimulusfassung für Gruppe B kam keines der vier Produkte als Product Placement im Spielfilm vor; die dazu entfernten Szenen wurden zu Werbespots umgeschnitten und zusammen mit anderen Werbespots in zwei Unterbrecherwerbeblöcken (anstelle der Product-Placment-Inseln) innerhalb von BIS ANS ENDE DER WELT gezeigt.

5.2 Ermittlung von Leistungsaspekt (d') und Antwortdisposition (B'') und Interpretation der Kennwerte

Um Leistungsaspekt (SDT: Diskriminationsleistung) d' und Antwortdisposition (SDT: Reaktionsneigung) B'' bestimmen zu können, wurden zunächst die Wahrscheinlichkeiten von ‚Treffer (H)' und ‚falscher Alarm (FA)' ermittelt und anhand dieser zwei Informationen der Leistungsaspekt d' (bzw. A_G) und die Antwortdisposition (B'') berechnet (siehe Abbildung 3). Für den nicht-parametrischen Fall des ja/nein-Recognition-Tests wird die Berechnung von AG nach Craig anstelle von d' vorgeschlagen (Shapiro, 1994).Für die Interpretation von AG und B'' gilt: Je größer AG, desto besser kann eine Person zwischen zuvor bereits gezeigten (‚alte Items') und in der Abfrageliste erstmals vorkommenden Stimuli (‚neue Items') unterscheiden. B'' kann Werte im Bereich von −1 bis +1 annehmen. Positive Werte von B'' verweisen auf ein konservatives Entscheidungsverhalten, d.h. eine Person, die fälschliche Wiedererkennungen (ein ‚neues' Items wird als ‚alt' bezeichnet = ‚falscher Alarm') zu vermeiden versucht und

[4] Merkmale der Stichprobe: N = 105; N_{Frauen} = 80; M_{Alter} = 21.4 Jahre.

[5] Merkmale der Stichprobe: N = 122 ; N_{Frauen} = 90; M_{Alter} = 24.8 Jahre.

$$A_G = \frac{p(H) + [1 - p(FA)]}{2} \qquad B'' = \frac{p(H) * [1 - p(H)] - p(FA) * [1 - p(FA)]}{p(H) * [1 - p(H)] + p(FA) * [1 - p(FA)]}$$

Abbildung 3: Formeln zur Berechnung von Leistungsaspekt (AG) und Antwortdisposition (B'')(Shapiro, 1994, S. 144).

damit riskiert, die Zahl der korrekten Wiedererkennungen (‚Treffer') zu reduzieren. Negative Werte von B'' verweisen auf ein progressives Entscheidungsverhalten: Personen die so raten, wollen möglichst viele ‚Treffer' erzielen, wodurch sie das Risiko für fälschliche Wiedererkennungen (‚falscher Alarm') erhöhen.

5.3 Ergebnisse

In Studie 1 zeigte sich kein Effekt der unterschiedlichen Programmintegration von Werbespots (siehe Tabelle 2): Die Wiedererkennungsraten unterschieden sich nicht, egal ob Werbespots für die sechs untersuchten Produkte/Marken innerhalb oder in Werbeblöcken zwischen Teilen der Unterhaltungsshow Die dicksten Dinger vorkamen. Nach der statistischen Analyse der Wiedererkennungsdaten zeigte sich selbiger Befunde für die Fähigkeit der Befragten, die Marken- oder Produktnamen im Fragebogen in alte (Marke- oder Produkt wurden zuvor im TV-Mitschnitt präsentiert) und neue Items (Marke- oder Produkt war nicht im TV-Mitschnitt enthalten) einzuteilen: In beiden Gruppen waren die Wiedererkennungsraten für die im Werbeblock bzw. in der Show gezeigten Zielmarken und -produkte (‚alte Reize') ebenso ausgeprägt wie AG als Maß für die Fähigkeit, zwischen diesen und den in der Abfrage des Wiedererkennungstests als Distraktoren präsentierten ‚neuen' Reizen unterscheiden zu können. Dieses Ergebnis legt nahe, beiden Darbietungsformen ein ähnliches Aktivierungs- und Wahrnehmungspotential zu unterstellen. Gemäß Signalentdeckungstheorie ist aber auch eine Interpretation derart vorstellbar, dass es tatsächlich eine Differenz im sensorisch Aktivierungspotential der Darbietungsformen Show oder Werbeblock gab, die zunächst unterschiedliche latente Antwortleistungen befördert haben könnte, in der manifesten Indikatorvariable ‚Wiedererkennungsraten' aber nicht sichtbar wurde, da in beiden Gruppen mit unterschiedlichen Dispositionen vorgegangen wurde. Die deskriptiven Werte für die Reaktionsneigungen deuten zunächst auf ein tendenziell progressiveres Entscheidungsverhalten (geringe Werte für B'') von Personen in der Sendungsgruppe im Vergleich zu Personen in der Werbungsgruppe hin. Allerdings sind die Unterschiede innerhalb der Gruppen bei Standardabweichungen von 0,73 (Sendungsgruppe) und 0,77 (Werbungsgruppe) höher als die Unterschiede zwischen den zwei Gruppen, so dass ein signifikanter Effekt der Darbietungsformen auf die Reaktionsneigungen auszuschließen ist.

Tabelle 2: Analyse des kognitiven Potentials von Informationsangeboten – ‚konventionelle' Auswertung des Wiedererkennungstests und Auswertung gemäß SDT für Studie 1 (N_{max} = 105).

Indikator	Produkt	Gruppenmittel: Sendungsgruppe	Gruppenmittel: Werbungsgruppe	*F*	*p*
		'üblicher' Wiedererkennungstest			
Wiedererkennungsrate[1]	GMX	.83	.85	.093	n.s
	IKEA	.98	.94	.991	n.s
	CONTINENTAL	.67	.68	.004	n.s
	TUI	.90	.94	.575	n.s
	SWATCH	.85	.79	.503	n.s
	TUBORG	.81	,81	.002	n.s
		Auswertung des Wiedererkennungstests gemäß SDT			
Diskriminationsleistung (A_G)		.37	.26	2.39	n.s
Reaktionsneigung (B'')		.20	.28	0.23	n.s

[1]Rate der korrekten Wiedererkennung [%]
[a]F[1,103]; [b]F[1,98]; [c]F[1,79]

Es ist also festzuhalten: Mit der Programmintegration von Werbespots wird das Wiedererkennungspotential der Marken- bzw. Produktinformationen weder verbessert noch verschlechtert – Werbespots in einem redaktionellen Programm statt in einem Werbeblock darzubieten ist unerheblich für die Wahrnehmung und das Wiedererkennen von Marken- und Produkten.

In Studie 2, dem Vergleich von Product Placements im Spielfilm BIS ANS ENDE DER WELT bzw. Werbespots im Werbeblock zwischen Teilen der Fernsehfassung dieses Spielfilms, fällt die Bewertung in Bezug auf die Wiedererkennungsraten dagegen anders aus als in Studie 1 (siehe Tabelle 3): Die Informationen zu einem Car-Info-Systems und einem Mobiltelefon von SONY sowie einem SHARP-Netbook wurden bei der Werbespotdarbietung aufmerksamer wahrgenommen und intensiver verarbeitet als bei einer Platzierung dieser Produkte im Spielfilm. Während knapp die Hälfte der Personen in Gruppe B (Werbespots) das SHARP-Netbook im Gedächtnistest korrekt wiedererkannte, war dies in Gruppe A (Product Placements) nur bei etwa jeder zehnten Person der Fall. Eine Ausnahme betrifft die Platzierung von LUFTHANSA im Spielfilm: Als sogenanntes Creative Placement war der LUFTHANSA-Airliner nicht nur Hintergrundkulisse, sondern selbst Handlungsträger und wurde deshalb ähnlich intensiv wahrgenommen wie bei der Werbespotdarbietung. Neben der besseren Gesamtwerte für die vier möglichen korrekten Wiedererkennungen (‚Wiedererkennungsrate gesamt') konnten Personen aus der Werbespotgruppe in der Wiedererkennungsaufgabe auch besser zwischen alten (Marke- oder Produkt wurde zuvor präsentiert) und neuen Reizen (Marke- oder Produkt war nicht enthalten) unterscheiden (‚AG') – ein Befund, der (abgesehen von dessen eingeschränkter Interpretierbarkeit aufgrund heterogener Fehlervarianzen) zunächst auf ein geringeres Aktivierungspotential von Product Placements im Vergleich zu Werbespots verweist. Wie zuvor diskutiert, lässt sich das Ergebnis des Wiedererkennungstests in Studie 2 aber auch ganz anders interpretieren: Wenn Personen in der ja/nein-Wiedererkennungsaufgabe zu Werbespotszenen einer anderen Antwortdisposition folgten als Personen in der ja/nein-Wiedererkennungsaufgabe zu Filmszenen, wäre die Darbietungsform der werblichen Botschaft in Bezug auf das Aktivierungspotential als ähnlicher einzuschätzen, auch wenn sich die Wiedererkennungsraten signifikant unterscheiden. Diese Annahme bestätigt sich aufgrund der Werte für die Reaktionsneigung: Personen in Gruppe A, also jene, die Marken und Produkte innerhalb des Spielfilms sahen, erweisen sich im Wiedererkennungstest eher als ‚konservative Rater'. Indem sie versuchten, fälschliche Wiedererkennungen (‚falscher Alarm') zu vermeiden, war eine im Vergleich zu den anderen Items in der Vorlagenliste des Wiedererkennungstests deutlich erhöhte Empfindungsstärke notwendig, einen Reiz als ‚zuvor gesehen' zu markieren. Die in den Filmszenen enthaltenen Marken-/Produktdarstellungen waren offenbaren nicht aufdringlich genug für eine elaborierte Verarbeitung und die Differenz zu einem ‚neuen' Item in der Abfrageliste zu gering – in dieser Unsicherheitssituation auf Nummer sicher gehend entschieden sich die Befragten im Wiedererkennungstest bei der überwiegenden Zahl von Items für die Antwort ‚zuvor nicht gesehen' und verpassten dadurch Items, für die eine Wiedererkennung korrekt (‚Treffer') gewesen wäre. Ganz anders gingen die Befragten in Gruppe B vor, denen Marken und Produkte per Werbespot dargeboten wurden: mit einem um fast 0,5 Punkte geringeren Wert für die Reaktionsneigung (bei einem Wertebereich von B'' von +1 bis -1) erweisen sich diese im Vergleich zu Personen in Gruppe A als deutlich ‚progressivere Rater', wenngleich B'' noch immer im konservativen Bereich und nahe dem theoretischen Mittel liegt.

Tabelle 3: Analyse des kognitiven Potentials von Informationsangeboten – ‚konventionelle' Auswertung des Wiedererkennungstests und Auswertung gemäß SDT für Studie 2 (N_{max} = 122).

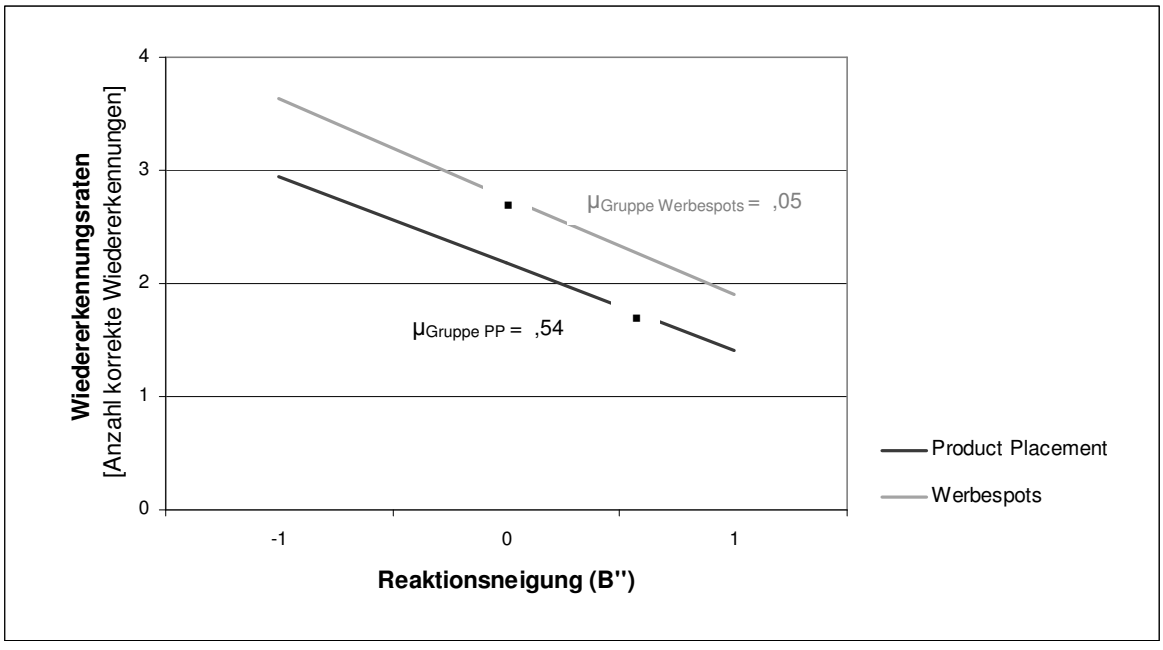
Indikator	Produkt	Gruppenmittel: Gruppe A: Product Placement	Gruppenmittel: Gruppe B: Werbespot	*F*	*p*
'konventioneller' Wiedererkennungstest					
Wiedererkennungsrate[1]	SONY (Car-Info-System)	.33	.55	6.13	.015
	LUFTHANSA	.79	.81	.071	n.s.
	SONY (Mobil-Telefon)	.26	.70	29.47	.000
	SHARP (Netbook)	.09	.52	32.76	.000
	gesamt	1.7	2.7	30.61	.000
Auswertung des Wiedererkennungstests gemäß SDT					
Diskriminationsleistung (A_G)[2]		.62	.74	13.10	.000
Reaktionsneigung (B'')[3]		.54	.05	15.87	.000

[1]Rate der korrekten Wiedererkennung [%]
[2]Test homegener Fehlervarianzen (Levene): L-Stat[1.110]=8.040; p = .005
[3]Test homegener Fehlervarianzen (Levene): L-Stat[1.110]=.033; p = .855

Indem Personen in Gruppe B versuchten, möglichst viele Treffer zu erzielen, d.h. ‚gute Werbungsentdecker zu sein', entschieden sie sich bereits bei einer (im Vergleich zu anderen Items in der Vorlagenliste des Wiedererkennungstests) minimal erhöhten Empfindungsstärke für die Antwortkategorie ‚zuvor gesehen' – auch wenn die Marken-/Produktinformationen in Gruppe B (Werbespots) gar nicht wesentlich aufdringlicher waren als jene in Gruppe A. Dadurch wurde zwar das Risiko eingegangen, einen Reiz als ‚zuvor gesehen' zu identifizieren, der im zuvor dargebotenen Werbeblock gar nicht vorkam – oft genug war diese Entscheidung aber (zufällig) richtig und die Wiedererkennung korrekt (ein ‚Treffer'). So entstand unter dem Eindruck der höheren Wiedererkennungsraten in Gruppe B der Eindruck, das Aktivierungspotential von Werbespots (mit Ausnahme LUFTHANSA) sei höher als das von Product Placements. Als Folge der separaten Betrachtung von Antwortleistung (Diskriminationsfähigkeit) und Antwortdisposition (Reaktionsneigung) in Studie 2 erscheint der Befund eines höheren Aktivierungspotentials von Werbespots gegenüber Product Placements nicht mehr ausreichend gültig.

Kritiker dieser Auffassung könnten einwenden, dass die als Reaktionsneigungen gemessenen unterschiedlichen Dispositionen, in einem Wiedererkennungstest bei Unsicherheit über die vorherige Wahrnehmung mit ‚zuvor gesehen' bzw. ‚zuvor nicht gesehen' zu antworten, eine Frage der Persönlichkeit, d.h. eine ‚Trait'-Eigenschaft sind und damit a priori vorhanden gewesen sein müssten. Faktisch können die beobachteten Gruppenunterschiede in den Reaktionsneigungen zum Wiedererkennungstest aber nur situative Realisationen dieser potentiell unterschiedlichen Antwortdispositionen in Folge der Darbietung unterschiedlicher Formate werblicher Botschaften sein – denn a) waren die interessierenden auditiven und visuellen Reizen in der Ursprungsstudie zur Reanalyse nahezu identisch, b) gab es im Unterschied zur üblichen Vorgehensweise bei Detektionsaufgaben (vgl. Fahr & Noessing, 2006) zuvor nicht die Instruktion, auf Werbespots bzw. Szenen im Spielfilm mit Produktplatzierungen zu achten und c) handelt es sich um ein kontrolliertes Experiment mit Randomisierung, bei dem alle Personen per Zufall auf die zwei Gruppen aufgeteilt wurden. Im Durchschnitt von A- und B-Gruppe waren die für einzelne Personen potentiell vielleicht unterschiedlich ausgeprägten Antwortendispositionen vor der Stimuluspräsentation damit ähnlich und können die Mittelwertunterschiede im Wiedererkennungstest (Wiedererkennungsraten und Diskriminationsleistungen) nach der Stimulusdarbietung nicht erklären (vgl. Rasch, Verdooren & Gowers, 1999).

Diese Interpretation belegt auch eine moderierte Regression zum Test der Interaktion der beiden angenommenen Verursachungsgrößen ‚Format der Informationsdarbietung' (Gruppe Product Placement versus Gruppe Werbespots) sowie ‚Reaktionsneigung' in der Vorhersage des manifesten Indikators ‚Wiedererkennungsraten'. Die Darbietungsform werblicher Botschaften und die Reaktionsneigung sind zunächst statistisch unabhängig (kein Interaktionseffekt; siehe Abbildung 4) in ihrem Zusammenhang zu den Leistungen im Wiedererkennungstest: Product Placements werden schlechter wiedererkannt als Werbespots und konservative Rater (höhere Werte von B'') erkennen schlechter wieder als progressive Rater (geringere Werte von B'').



Test der Effekte im Modell $y = b_0+b_1B''+b_2Stimulus+b_3B'' \times Stimulus$:
- B'' (Reaktionsneigung): b = -.766; t = -3.775; p =.000
- Stimulus (PP versus Werbespots): b = .587; t = 3.153, p = .002
- B'' x Stimulus: b =-.104; t = -0.425, p = .67

Abbildung 4: Test des Interaktionseffekts von ‚Format der Informationsdarbietung' und ‚Antwortneigung' in der Vorhersage von Wiedererkennungsrate.

Betrachtet man jedoch die bedingten Effekte wird deutlich, dass sich der Effekt der Darbietungsform des Informationsangebotes mit Veränderungen der Antwortdisposition in Richtung ‚konservatives Raten' tendenziell verringert: Werbespots sind in Studie 2 zwar auch dann noch Product Placements überlegen, wenn Befragte statt viele Treffer (Items, die zuvor gezeigt waren, als ‚wiedererkannt' markieren) zu erzielen lieber weniger Fehler (im Sinne von falscher Alarm, d.h. Items als ‚wiedererkannt markieren', die zuvor nicht gezeigt waren) machen wollen (höhere Werte von B'') – im direkten Stichprobenvergleich, d.h. an den Stellen, die die durchschnittlichen Ausprägungen von Reaktionsneigungen in den jeweiligen Stimulusbedingungen markieren (Gruppe mit Werbespots = 0,05; Gruppe mit Product Placements = 0.54), ist die Differenz in den Wiedererkennungsraten mit 0.97 aber deutlich größer als der Unterschied, der für zwei Personen mit gleichen Reaktionsneigungen (z. B. 0,59 bei B''=0) feststellbar ist – faktisch ist der Unterschied in den kognitiven Potentialen der beiden Informationsdarbietungen Werbespot und Product Placement geringer als über den manifesten) Indikator vorhergesagt.

Mit der moderierten Regression wurde zudem die kritische Frage untersucht, ob man Größen wie ‚Diskriminationsleistung' und ‚Reaktionsneigung' als Kontrollvariablen im Hinblick auf eine Kriteriumsvariable verwenden kann, wenn diese aus der manifesten Indikatorvariable abgeleitet sind. Sowohl bei der ‚Reaktionsneigung' als auch beim Interaktionsterm ‚Format der Informationsdarbietung' x ‚Reaktionsneigung') wurden die Grenzen im Test der Multikollinearität (zum Prädiktor ‚Format der Informationsdarbietung' als angenommenem Haupteffekt für die Wiedererkennungsraten) nicht überschritten (‚Reaktionsneigung': Toleranz = 0.270; VIF = 3.709; ‚Format der Informationsdarbietung' x ‚Reaktionsneigung': Toleranz = 0.312; VIF = 3.2005). Damit lässt sich auch ausschließen, dass die Strategie, Wiedererkennungsraten in einen Leistungsaspekt und in eine Antwortdisposition aufzuteilen, um so eine Kontrolle des manifesten Indikators für den Medieneffekt zu erreichen, dadurch konterkariert wird, dass die Kontrollgröße selbst konfundiert ist.

6 Fazit der (Re)Analyse – Relevanz der Unterscheidung von Leistungsaspekt und Antwortdisposition in medien- und werbepsychologischen Studien

Der vorliegende Beitrag untersuchte, inwiefern die in Latent-Trait-Analysen (LTA) bzw. in der Item-Response-Theorie (IRT) übliche Aufteilung manifester Indikatoren in eine Aufgabenschwierigkeit und in eine Personenfähigkeit auf Messungen zum Gedächtnis übertragen werden kann. Wie gezeigt liefert die Signal-Entdeckungs-Theorie (SDT) einen geeigneten Ansatz, um Wiedererkennungsdaten analog zur IRT in zwei Informationen aufzuteilen:

In eine mit der (latenten) Personenfähigkeit vergleichbare Diskriminationsleistung (‚alte' von ‚neuen' Reizen unterscheiden zu können) sowie eine mit der Aufgabenschwierigkeit vergleichbare Neigung, bei Unsicherheit über die Vertrautheit mit ‚alten' bzw. ‚neuen' Reizen eher die Antwortkategorie ‚zuvor gesehen' oder eher die Antwortkategorie ‚zuvor nicht gesehen' zu wählen.

Auf Recall-Tests lässt sich das Modell der SDT prinzipiell ebenso anwenden; allerdings fehlt hier die Möglichkeit, analog zum ‚falschen Alarm' bzw. zur ‚korrekten Zurückweisung' in Wiedererkennungstests die nicht korrekte Erinnerung bzw. korrekte Nicht-Erinnerung zuvor nicht dargebotener Informationen über Kategorien mit prinzipiell gleicher Verteilungswahrscheinlichkeit aufzuzeichnen.

Die Erläuterungen zum Informationswert von Auswertungen gemäß SDT verdeutlichen den Erkenntnisgewinn, den die Anwendung des Konzepts lokaler stochastischer Unabhängigkeit grundsätzlich erbringt: Sie zeigen, dass es potentiell zwei und nicht nur eine Ursache für hohe Wiedererkennungsraten, d.h. für eine hohe Anzahl von korrekt wiedererkannten Reizen gibt: a) Medienangebote weisen tatsächlich ein hohes kognitives Potential auf (erhöhen Aufmerksamkeit, leiten das Lernen neuer bzw. den Abruf bekannter Konzepte an) und/oder b) Personen sind ‚progressive Rater', d.h. sie markieren Items in Wiedererkennungstest bereits dann als zuvor gesehen, sobald sie eine minimale Vertrautheit mit entsprechenden Reizen empfinden und ohne sich sicher zu sein, dass ihre ‚Wiedererkennungen' korrekt sind.

Während diese Erkenntnis für die mit Beobachtungsdaten arbeitende Medienpraxis im Unterschied zur Häufigkeit von Analysen nach der SDT schon immer hoch relevant war, erschien die Fokussierung auf Wiedererkennungsraten in experimentellen Studien bisher kein Problem. Unter der Annahme, die Antwortdisposition weise ähnlich wie Trait-Variablen in der Persönlichkeitspsychologie eine zeitliche und situative Konsistenz auf und sei mit der manifesten Antwortleistung unkorreliert, galt das Experiment als Methode der Wahl, potentielle Dispositionsunterschiede zu kontrollieren. Die Reanalyse von zwei experimentellen Studien mit ja/nein-Wiedererkennungsaufgaben zeigt aber, dass sich diese Annahme empirisch nicht halten lässt: Antwortdispositionen sind zumindest teilweise situativ, denn sie unterschieden sich trotz randomisierter Gruppenbildung, identischen Instruktionen aber unterschiedlichen Formaten der Darbietung werblicher Botschaften im Gruppenvergleich nach der Stimuluspräsentation. Wäre die Antwortdisposition ein ‚Trait' und stochastisch unabhängig von der latenten Fähigkeit einer Person (einen konkreten Reiz wiederzuerkennen), hätte die Randomisierung vorab bestehende individuelle Unterschiede (z. B. Werbebotschaften leicht zu entdecken wegen höherer Coping-Kompetenz; ‚gesehen zu sagen' egal ob man Fehler macht) egalisiert, sodass sie faktisch zwar immer noch die Ausprägung der manifesten Indikatorvariable bestimmen, das Basislevel der Wiedererkennungsraten aber synchron und ohne Unterscheide zwischen den Gruppen festlegen.

Forschungsdesigns, die Drittvariableneinflüsse durch Zufallsverteilung von Personen auf Gruppen oder mehrere Messungen an derselben Person zu kontrollieren versuchen, können das Problem der Mehrdeutigkeit von Wiedererkennungsraten ebenso wie Kovarianz- und Interaktionseffektanalysen aber nicht lösen, wenn die manifeste Antwortleistung (‚Wiedererkennungsraten') und die Antwortdisposition (‚Reaktionsneigung') eine gemeinsame Ursache haben, d.h. im Test konvergieren und nicht vollständig unabhängig voneinander erhoben werden können.

Unterschiedliche Antwortdispositionen, die aus der Testaufgabe oder Hinweisen im Studienverlauf (z. B. die Art des Informationsangebotes) resultieren, dürften nicht nur Kennzeichen und untrennbarer Bestandteil des hier behandelten Wiedererkennungstests sein. Zwar erwiesen sich die Wiedererkennungsraten in der diskutierten Studie 2 mit wenigen Product Placements bzw. Werbespots im Informationsangebot und wenigen Items in der Vorlage zur ja/nein-Wiedererkennungsaufgabe als Indikator für Medieneffekte, den Diskriminationsleistungen und Antwortneigungen zwar relativieren, aber nicht widerlegen – bei anderen Informationsdarbietungen bzw. Testkonstellationen sind konträre Ergebnisse zu erwarten.

Wie eingangs beschrieben wurde, ist daher kritisch zu fragen, inwiefern die bisher weitgehende Nichtbeachtung der situativen Abhängigkeit der Antwortneigung von Darbietungs- und Testbedingungen als unproblematisch gelten kann. Das trifft insbesondere für Studien im Kontext globalisierter Marken- und Produktkommunikation zu, wo zu fragen ist, ob beobachtete Länderunterschiede in den Urteilen von Konsumenten tatsächlich eine substantielle Differenz anzeigen, oder nur auf unterschiedlichen Antwortdispositionen beruhen, für die kulturelle Unterschiede zwischen Ländern (vgl. Hofstede, 2001) maßgeblich sind. So gesehen ist die Annahme zielführend, dass eine Reihe von manifesten Indikatoren in medien- und werbepsychologischen Studien nicht nur Erkenntnisse über das persuasive oder lernbezogene Potential von Informationsangeboten (seien es Angebote im TV, Online-Medien, Büchern oder der interpersonalen Kommunikation) enthalten, sondern auch Ergebnis einer bestimmten Antwortdisposition sind. Antwortdispositionsunterschiede anzunehmen heißt nicht, jeden Test und jede Befragung zu problematisieren: man sollte die Bedingung der Möglichkeit des Auftretens aber bedenken und im Sinne eines ‚Differential Item Functioning' (vgl. Embretson & Reise, 2000) nicht nur in Beobachtungsstudien, sondern auch in experimentellen Studien prüfen, um Effekte von Informationsangeboten zur Gesellschafts- und Wirtschaftskommunikation angemessen bestimmen zu können.

7 Literaturverzeichnis

Cohen, J., Cohen, P., West. S. G. & Aiken, L. S. (2002). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Hillsdale: LEA.

Dardis, F. E., Schmierbach, M. & Limperos, A. M. (2012). The Impact of Game Customization and Control Mechanisms on Recall of Integral and Peripheral Brand Placements in Videogames. *Journal of Interactive Advertising, 12*, 1-12.

Embretson, S. E. & Reise, S. P. (2000). *Item Response Theory for Psychologists*. New Jersey: LEA.

Engelhardt, A. (1999). *Werbewirkungsmessung : Hintergründe, Methoden, Möglichkeiten und Grenzen*. München: Verlag R. Fischer.

Fahr, A. & Noessing, B. (2006). *Werbung im Programm. Wie grenzen Zuschauer Werbung, Product Placement und Sponsoring während der Rezeption ab?* Vortrag anlässlich der Münchner Medientage, München, 20. Oktober 2006.

Hofstede, G. (2001). *Culture's Consequences – Comparing Values, Behaviors, Institutions and Organizations Across Nations*. Thousand Oaks: Sage.

Klein, B. (2009). *Zur Werbewirksamkeit von In-Game-Advertising*. München: FGM Verlag.

Kolb, S. & Beck, D. (2011). Vergleichbarkeit in der (international) vergleichenden Journalismusforschung auf der Basis von Sekundäranalysen. In T. Quandt & O. Jandura (Hrsg.). *Methoden der Journalismusforschung* (S. 351-366). Wiesbaden: Verlag für Sozialwissenschaften.

Kubinger, K. D. (2005). Psychological Test Calibration using the Rasch Model - Some Critical Suggestions on Traditional Approaches. *International Journal of Testing, 5,* 377-394.

Mackay, T., Ewing, M. Newton, F. & Windisch, L. (2009). The effect of product placement in computer games on brand attitude and recall. *International Journal of Advertising, 28,* 423–438.

Mokken, R. J. (1997). Nonparametric models for dichotomous responses. In R. K. Hambleton & W. J. Van der Linden (Hrsg.). *Handbook of Modern Item Response Theory* (P. 351-367). New York-Berlin: Springer-Verlag.

Neubauer, A. C. & Knorr, E. (1998). Three paper-and-pencil tests for speed of information processing: Psychometric properties and correlations with intelligence. *Intelligence, 26*, 123–151.

Neyman, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. *Statistical Science, 5*, 465–480.

Nicovich, S. G. (2005). The Effect of Involvement on Ad Judgment in a Video Game Environment: The Mediating Role of Presence. *Journal of Interactive Advertising, 6*, 29-39.

OQ1: Eurobarometer (2010). *Science and Technology Report EB73.1*. http://ec.europa.eu/public_opinion/archives/ebs/ebs_340_en.pdf. (abgerufen am 02.07.2012).

Partchev, I. & De Boeck, P. (2012). Can fast and slow intelligence be differentiated? *Intelligence, 40*, 23–32.

Rasch, D., Verdooren, L. R. & Gowers, J. I. (1999). *Fundamentals in the Design and Analysis of Experiments and Surveys.* München und Wien: Oldenburg.

Richter, T. (2007). Wie analysiert man Interaktionen von metrischen und kategorialen Prädiktoren? Nicht mit Median-Splits! *Zeitschrift für Medienpsychologie, 19,* 116-125.

Russel, C. A. (2002). Investigating the effectiveness of product placement in television shows: The role of modality and plot connection congruence on brand memory and attitude. *Journal of Consumer Research, 29,* 306-318.

Rypma, B. & Prabhakaran, V. (2009). When less is more and when more is more: The mediating roles of capacity and speed in brain-behavior efficiency. *Intelligence, 37,* 207–222.

Scheiblechner, H. (2007). A unified Nonparametric IRT Model for d-Dimensional Psychological Test Data (d-ISOP). *Psychometrika, 72*, 43–67.

Schemer, C. (2007). Wem Medienschönheiten schaden: Die differenzielle Anfälligkeit für negative Wirkungen attraktiver Werbemodels auf das Körperbild junger Frauen. *Zeitschrift für Medienpsychologie, 19,* 58-67.

Schmitt, M. (2004). Persönlichkeitspsychologische Grundlagen. In Mangold, R.; Vorderer, P. & Bente, G. (Hrsg.). Lehrbuch der Medienpsychologie (151-173). Göttingen: Hogrefe.

Schmitt, M., Gollwitzer, M., Baumert, A., Gschwendner, T., Hofmann, W. & Rothmund, T (2008). *Traits as Situational Sensitivities. Psychometric and Substantive Comments on the TASS Model Proposed by Marshall and Brown (2006).* Bericht aus der Arbeitsgruppe "Verantwortung, Gerechtigkeit, Moral" der Universität Koblenz-Landau, Fachbereich Psychologie. http://psydok.sulb.uni-saarland.de/volltexte/2009/2352/.

Shapiro, M. E. (1994). Signal detection measures of recognition memory. In A. Lang (Hrsg*.). Measuring psychological responses to media* (P. 133-148). Hillsdale: LEA.

Sheppard, L. D. & Vernon, P. A. (2007). Intelligence and speed of information processing: A review of 50 years of research. *Personality and Individual Differences, 44*, 247–259.

Slater, M. D., Hayes, A. F., Reineke, J. B., Long, M. & Bettinghaus, E. P. (2009). Newspaper Coverage of Cancer Prevention: Multilevel Evidence for Knowledge-Gap Effects. *Journal of Communication, 59,* 514–533.

Steyer, R. & Eid, M. (2001). *Messen und Testen*. Berlin: Springer.

Steyer, R., Partchev, I., Kroehne, U., Nagengast, B., & Fiege, C. (2010). *Probability and Causality: Theory. Manuscript in preparation*. Heidelberg: Springer. http://www.metheval.uni-jena.de/get.php?f=1032.

Suedfeld, P. & Tetlock, P. E. (2003). Individual differences in Information processing". In A. Tesser & N. Schwarz (Hrsg.). *Intraindividual processes* (P. 284-304). Malden: Blackwell.

Taylor, D. G., Strutton, D. & Thompson, K. (2012). Self-enhancement as a motivation for sharing online advertising. *Journal of Interactive Advertising, 12*, 13-28.

Velden, M. (1982). *Die Signalentdeckungstheorie in der Psychologie.* Stuttgart: Kohlhammer.

Wirth, W. & Kolb, S. (1999). *Wissensmessung in der Kommunikationsforschung. Vortrag auf der 1. Tagung der DGPuK-Fachgruppe Methoden*, Leipzig 1999.

Wirth, W., Heydecker, A. & Kolb, S. (2006). *Wissensmessung in der Kommunikationswissenschaft. Eine Vollerhebung der Konzeptualisierung und Operationalisierung von Wissen in zwölf nationalen und internationalen Fachzeitschriften von 1970-2005*. Vortrag auf der 8. Tagung der DGPuK-Fachgruppe Methoden, Zürich.

Woelke, J. (1998). Product Placements oder Werbespots? Zwei Werbepräsentationsformen im Vergleich. *Zeitschrift für Sozial-psychologie, 29*, 165-174.

Woelke, J. (2004). *Durch Rezeption zur Werbung. Kommunikative Abgrenzung von Fernsehgattungen*. Köln: Halem Verlag.

Woelke, J. (2008). Nicht alle, aber einige mehr. Werbewirkungen unter dynamisch-transaktionaler Perspektive. In C. Wünsch, W. Früh, W. & V. Gehrau (Hrsg). *Integrative Modelle in der Rezeptions- und Wirkungsforschung. Dynamische und transaktionale Perspektiven* (S. 81-106). München: Fischer Verlag.

Woelke, J. & Dürager, A. (2011). Interpersonale Beeinflussbarkeit und Werbewirkung. Erstellung einer deutschsprachigen Version der ‚CSII' und Test für die Medien- und Werbepsychologie. *Journal of Business and Media Psychology, 2*, 1-9.

Korrespondenzadresse:

Dr. Jens Woelke
Empirische Kommunikations- und Medienforschung
Universität Leipzig
Burgstraße 21
D-04109 Leipzig
GERMANY

jens.woelke@uni-leipzig.de