Copytests: Zwei Abfragen, - zwei Ergebnisse: Was nun ?

Diese Analyse befasst sich mit der Frage, was geschieht, wenn man die „Codierung“ von Antworten auf offene Fragen den Befragten überlässt. Obwohl das Beispiel aus der Werbeforschung stammt, ist es von allgemeiner Bedeutung

Zur Anzeigen-Beachtung

Für Anzeigen, Fernsehspots und alle anderen Werbemittel gilt die Binsenweisheit, dass die Beachtung die notwendige (aber nicht ausreichende) Voraussetzung dafür ist, irgendeine Wirkung ausüben zu können.

Das klassische Verfahren, um die Beachtung zu ermitteln, ist das Verfahren des Wiedererkennens, mit dem englischen Begriff „Recognition“ auch in Deutschland bezeichnet. Dieses Verfahren ist insbesondere seit den 60er Jahren stark angezweifelt worden, ausgehend von einem Aufsatz von Prof. Lucas. Koeppler hat darüber berichtet. Das Verfahren und die deutsche Historie habe ich an anderer Stelle kurz beschrieben.

Diese Kritik muss uns hier aus zwei Gründen nicht beschäftigen: Erstens geht es nachfolgend nicht um die Methode an sich, sondern um ein Detailproblem, das unabhängig von der grundsätzlichen Frage der Brauchbarkeit der „Recognition“-Methode auftritt.

Zweitens habe ich kürzlich eine umfassende Prüfung der „Recognition“-Kritik abgeschlossen, deren Ergebnis ist: Die Kritik ging einerseits von falschen Voraussetzungen aus und beruht andererseits auf einem falsch konzipierten Test.

Zwei Verfahren im Vergleich

Ausgangspunkt der nachfolgenden Analyse waren die Ergebnisse zweier Copy-Tests, die in „interview und analyse“ (Februar 1983) veröffentlicht wurden. Die Quelle hierfür war eine Publikation der Gruner & Jahr-Marktforschung, die sich auf Tests des Instituts Media Markt Analysen für den „stern“ und eine andere Zeitschrift bezog[1].

Bei diesen Tests wurden etwas unterschiedliche Verfahren verwendet, die bei nahezu identischen Anzeigen unterschiedliche Resultate lieferten. Deshalb ist ein Vergleich der Ergebnisse von Zeitschrift zu Zeitschrift unmöglich. Damit könnte man die Betrachtung abschließen und sagen: Sowas kommt von sowas!

Es lohnt sich jedoch aus grundsätzlichen Erwägungen, die Daten einer weitergehenden Analyse zu unterziehen.

[1] Dr. Eva Maria Hess und Jörg Laufer lieferten freundlicherweise die erforderlichen Daten und Auskünfte

Die zwei Verfahren

Zunächst eine Beschreibung der zwei Verfahrensweisen:

In beiden Fällen wendet man sich an Leser(innen) der jeweiligen Publikation, welche die betr. Ausgabe bereits durchgeblättert oder gelesen haben. Mit ihr/ihm geht die/der Interviewer(in) das Heft durch und fragt zu den für den Test in Betracht gezogenen redaktionellen Beiträgen und Anzeigen, ob diese gesehen worden sind.

Zu dem einzelnen Beitrag wird abgefragt, ob er ganz, teilweise gelesen, nur die Überschrift gelesen bzw. Bilder angesehen wurden, - ob nichts gelesen bzw. angesehen wurde. Zur einzelnen Anzeige wird nur nach „gesehen" oder „nicht gesehen" gefragt.

Von hier an wird es unterschiedlich:

1.) „Anstreichen":
Zu jeder als „gesehen" bezeichneten Anzeige wurde gefragt:

> *„Zeigen Sie mir bitte jetzt alles, was Sie damals auf dieser Anzeige gesehen oder gelesen haben. Haben Sie das Markenzeichen (den Firmennamen) gesehen? Zeigen Sie mir jetzt bitte, wieviel Sie vom Text gelesen haben. Was haben Sie sonst noch auf dieser Anzeige bemerkt?"*
>
> (Im Testheft streicht die/der Interviewer(in) alles an, was die/der Befragte als „gesehen" oder „gelesen" genannt hat.)

Die Kategorisierung und Verschlüsselung der durchgestrichenen Elemente erfolgt im Institut.

2.) „Fragebogen-Eintragung":

Zu jeder als „gesehen" bezeichneten Anzeige wurde gefragt:

> „*Haben Sie den Text damals auf dieser Anzeige ganz gelesen, mehr als die Hälfte gelesen, weniger als die Hälfte gelesen oder gar nicht gelesen?"*
>
> *„Haben Sie damals den Marken-/Firmennamen bemerkt oder nicht bemerkt?"*
>
> (Die/der Interviewer(in) hat die entsprechenden Antwortvorgaben im Fragebogen zu kringeln).

Ergebnisausweis

Entsprechend den von Dr. Daniel Starch eingeführten Kategorien wurden drei Ergebnisse ausgewiesen:

1) Anzeige gesehen
2) Die Hälfte oder mehr vom Text gelesen
3) Marken-/Firmennamen gelesen/bemerkt.

Der erste Messwert reflektiert die Beachtung von mindestens einem Detail der Anzeige, ist also weitgehend bestimmt von auffälligen Elementen wie Illustration, fetter Überschrift oder Slogan.

Der Tabelle kann man die z.T. großen Unterschiede entnehmen, mit gegenläufigen Tendenzen für zwei der Kategorien:

- Die Fragebogen-Abfrage liefert eine höhere Mindestbeachtung;
- das Anstreichen liefert höhere Werte für das intensivere Lesen;
- die Beachtung von Marke/Hersteller zeigt keine generelle Schlagseite für die eine oder die andere Methode.

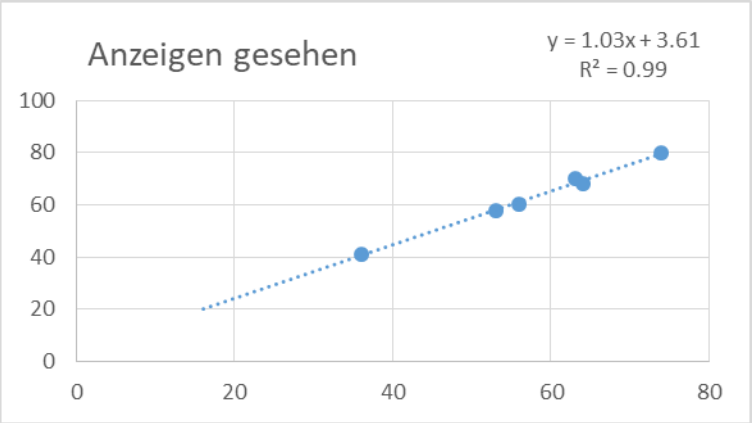
Tabelle: Vergleich der Ergebnisse

Anzeige für ...	Anstreichen %	FB-Eintragung %
Jägermeister		
Anzeige gesehen	74	80
Marke/Hersteller beachtet	74	75
1/2 und mehr vom Text gelesen	66	36
Lift		
Anzeige gesehen	64	68
Marke/Hersteller beachtet	63	55
1/2 und mehr vom Text gelesen	25	17
Camel		
Anzeige gesehen	63	70
Marke/Hersteller beachtet	61	63
1/2 und mehr vom Text gelesen	42	14
B H W		
Anzeige gesehen	53	58
Marke/Hersteller beachtet	41	44
1/2 und mehr vom Text gelesen	13	9
Longines		
Anzeige gesehen	36	41
Marke/Hersteller beachtet	36	34
1/2 und mehr vom Text gelesen	13	11
Lancia		
Anzeige gesehen	56	60
Marke/Hersteller beachtet	56	52
1/2 und mehr vom Text gelesen	25	26

Weiterführende Analyse

Als nächstes möchte man wissen, ob die Ergebnisse der beiden Verfahrensweisen irgendwelche Beziehungen zueinander haben. Deshalb untersuchen wir nun die Messwerte-Paare mit Hilfe von Regressionsanalysen.

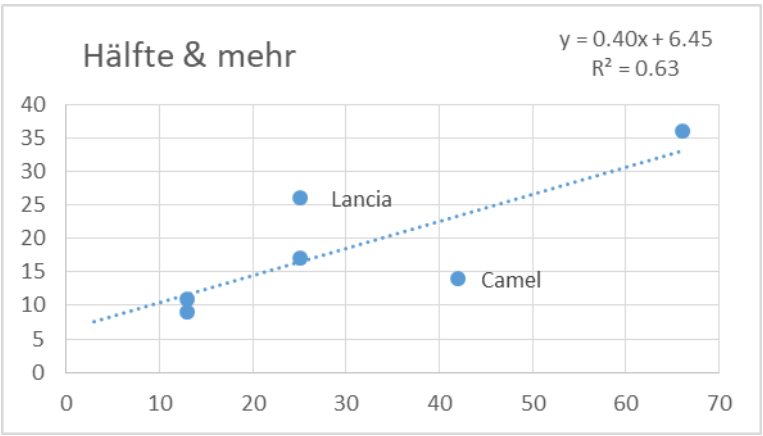
a) „Anzeige gesehen“
Die Ergebnisse der beiden Verfahren korrelieren extrem hoch (r = .996, erkl. Varianz: 99,2%); die erste Grafik veranschaulicht dies. Es gibt keine ins Gewicht fallenden Abweichungen der Messwerte von den Erwartungswerten:



Das heißt: Bezogen auf ein Mindestmaß der Beachtung, spielt die Methode für einen Vergleich der Anzeigen untereinander keine Rolle; es gibt nur eine Niveau-Verschiebung.

b) „Marke/Hersteller beachtet“
Die Unterschiede sind in 5 der 6 Fälle gering und gehen in beide Richtungen; es existiert keine eindeutige Tendenz und dementsprechend keine Korrelation. Auf den Unterschied von 8% bei der „Lift“-Anzeige ist noch einzugehen.

c) „Die Hälfte und mehr gelesen“
Hierbei sind die Werteniveaus gegenüber der globalen Beachtung vertauscht: die Anstreich-Methode liefert bei 4 Anzeigen deutlich höhere Prozente. Die Korrelation ist geringer, bei 2 Anzeigen weichen die Messwerte von den Erwartungswerten stark ab.



Daraus wäre zu schließen, dass die Anzeigengestaltung von 4 Anzeigen ziemlich ähnlich ist, wohingegen die beiden Ausreißer in gegensätzlicher Weise davon abweichen.

Das ist in der Tat so:

- Die Anzeige für „**Camel**“ enthält gar keinen „normalen“ Text, sondern nur einen Slogan;
- in der „**Lancia**“-Anzeige steht überdurchschnittlich viel Text.

Jedoch ist zu konstatieren, dass die Regressionsanalyse in diesem speziellen Fall das falsche Analyseinstrument ist: Ihre Ergebnisse verleiten zur Annahme, dass bei hohen Beachtungswerten ein großer Unterschied zwischen den beiden Methoden zu „erwarten“ sei, bei niedrigen Werten wenig oder gar nicht. Dafür dürfte es keine Begründung geben.

Stattdessen müssen wir die absoluten Unterschiede zugrunde legen und die Anzeigen danach sortieren.

Bei den Anzeigen für „Jägermeister“, „Camel“ und „Lift“ gibt es die größten Prozent-Unterschiede. Diese Anzeigen enthalten wenig Text.

Die kurzen Texte in den Anzeigen für „Jägermeister“ und „Lift“ sind so geartet, dass der Betrachter Schwierigkeiten hat abzuschätzen, was davon die Hälfte ist. Die Anzeige für „Camel“ enthält keinen „normalen“ Text, sondern nur einen Slogan sowie die übliche Fußnote zur Schädlichkeit des Rauchens. Was ist davon „die Hälfte“ des Textes?

Im Kontrast dazu sind die Texte bei den Anzeigen für „BHW“, „Longines“ und „Lancia“ länger; also kann man eher entscheiden, was davon die Hälfte sein könnte.

Erörterung der Resultate

Die Unterschiede der beiden Ermittlungsverfahren haben unterschiedliche Gründe, weshalb sie für die Messwerte getrennt behandelt werden müssen.

<u>Mindestmaß an Beachtung</u>

Die Vorfrage nach der Beachtung der redaktionellen Beiträge und die Frage danach, ob die betr. Anzeige „gesehen“ wurde, sind in beiden Versionen identisch. Für die durchgängig niedrigeren Werte der Anstreich-Methode oder die höheren der Abfrage-Methode ist von daher also keine Begründung zu finden.

Es bleiben zwei Möglichkeiten offen, soweit ich es beurteilen kann:

1. Die jeweils andersgeartete nachfolgende Ermittlung hat bei Befragten und/oder Interviewern zu geringfügig unterschiedlichen Lernvorgängen geführt. Das könnte z. B. die Vorstellung sein, dass unter „gesehen“ weniger oder mehr verstanden wird. Und bei der Abfragerei/ dem Anstreichen könnte es sein, dass die größere Mühsal sich auf die Antworten auswirkt.

   Ob es irgendeine solche Tendenz gibt, ließe sich nur dadurch überprüfen, dass

man die Ergebnisse aus einer größeren Anzahl von Tests auf Positionseffekte im Interview prüft. Das ist wegen der geringen Zahl von Anzeigen mit dem vorliegenden Material nicht möglich.

2. Es gibt Unterschiede zwischen den Zeitschriften bzw. deren Leserschaften. Das könnte jedoch nur mit identisch angelegten Tests in den zwei Zeitschriften überprüft werden.

Marke/Hersteller bemerkt

Es gab einen größeren Unterschied und zwar bei der „Lift"-Anzeige zugunsten des Anstreichens.

Hier könnten zwei Definitionsprobleme aufgetreten sein. Der Name „Lift" stand in der Überschrift, auf einer Flasche und auf 4 aufgereihten Kronenkorken.

Haben alle Befragten alle diese Nennungen bei der Frage berücksichtigt, ob sie den Markennamen „bemerkt" hatten? Oder haben einige gedacht, es sei z.B. jener in der Überschrift nicht gemeint, weil er nicht in der markentypischen Schrift geschrieben war?

Beim Anstreichen wurde zwar separat nach dem „gesehen" des Markenzeichens gefragt, und da könnte ggf. dasselbe Problem aufgetreten sein; aber dann wurde nach allen anderen Details gefragt, was sonst noch beachtet worden war. Ob darunter der Markenname war, wurde im Institut erkannt und ggf. zusätzlich verschlüsselt.

Als zweites findet man zwei Unterschiede in der Fragestellung:

- Beim Anstreichen: „ ... das Marken*zeichen gesehen*?"
- Beim Abfragen: „ ... den Marken*namen bemerkt*?"

Dies sind deutlich qualitative Unterschiede. Sie können die Ergebnisse beeinflusst haben; aber ich wage nicht, eine Tendenz anzugeben. Man erkennt, wie genau man sich bei der Fragenformulierung überlegen muss, was man wissen möchte!

Die Hälfte und mehr gelesen

Grundsätzlich soll dieser Messwert ein Indiz für die intensivere Beschäftigung mit den verbalen Inhalten der einzelnen Anzeige liefern. Das ist zu bedenken, wenn man ihn einsetzt.

Die stark unterschiedlichen Ergebnisse für diese Kategorie muss man wohl als „Instrumenten-Fehler" bezeichnen. Mit dieser Feststellung übe ich auch Selbstkritik, denn dergleichen habe ich häufiger verwendet. Das Ausmaß des Problems ist mir erst durch diese Analyse bewusst geworden.

Wenn man eine Kategorie etabliert, wie „die Hälfte oder mehr vom Text beachtet", muss man zwangsläufig definieren, was zum „Text" gehören soll, und wie man die Menge misst.

Bei der Betrachtung unterschiedlicher Anzeigen daraufhin, was als „Text“ anzusehen ist, gerät man alsbald in Schwierigkeiten:

- Gehört der kurze Slogan zum Text?
- Der Name der Marke bzw. des Herstellers?
- Die Bildunterschrift?
- Der Text auf der abgebildeten Packung?
- Die Fußnote (z.B. bei Zigaretten)?

Das Problem wird durch die Frage verschärft, welche Maßeinheit zu nehmen ist:

- Spalten-cm?
- Quadratzentimeter (also Flächen)?
- Anzahl der Textelemente?
- Oder Anzahl der Worte?

Dahinter steht die Frage, ob der fett gedruckte Text so viel wert ist wie die gebrauchte Fläche oder die Anzahl der Worte, – im Vergleich zu kleiner gedrucktem Text?

Mit wenig Phantasie ist das Dilemma auszumalen, in das man gerät.

Dies ist keine akademische Betrachtung, denn es darf nicht vergessen werden, dass jegliche Art des Vermessens die Ergebnisse und damit die Vergleiche beeinflusst! Unweigerlich gerät man in die Situation, Unterschiede zu erhalten und erklären zu müssen, die zum Teil durch objektive Unterschiede der Anzeigen, zum anderen Teil aber durch die gewählte Maßeinheit verursacht worden sind. Inserenten und Anzeigengestalter finden derartige Interpretationen wenig aufschlussreich und überzeugend. Als Forscher kommt man der berühmten (leider nicht näher bezeichneten) Marktfrau nahe, die den Finger auf die Waage legt und fragt: „Darf's ein wenig mehr sein?“

So viel zur Schwierigkeit, innerhalb des Instituts eine auch nur halbwegs brauchbare Definition des Textes und seiner Vermessung zu finden.

Bei dem zweiten Verfahren soll die befragte Person sagen, ob sie die Hälfte oder mehr vom Text beachtet habe. Aber was ist ihre Definition von Text? Und was ist ihre Vorstellung, welche Maßeinheit sie nehmen sollte?
Es ist davon auszugehen, dass sie es nicht weiß,– es gar nicht wissen kann. Überdies hat jede befragte Person ihre eigene Vorstellung. Das jedoch bedeutet: Die/der Mediaforscher (in) weiß nicht, was die Antworten aussagen.

<u>Genau das spiegelt sich in den Ergebnissen wider:</u>

Die Anzeigen mit einem erkennbaren „normalen“ Text kommen bei beiden Methoden ungefähr gleich weg: die Unterschiede betragen 1% (Lancia), 2% (Longines) und 4% (BHW). Hier fanden Mediaforscher und Befragte anscheinend nahezu gleichartige Abschätzungen.

Bei den anderen Anzeigen, wo man nicht sagen kann, was denn mit „Text“ gemeint ist, betragen die Unterschiede 8% (Lift) und 30% (Jägermeister); bei der Anzeige ohne jeglichen „normalen“ Text (Camel) 28%.

Mediaforscher und Befragte kommen in diesen Fällen zu ganz unterschiedlichen Definitionen – wen wundert's?

Dies führt zum Schluss, dass insbesondere für Anzeigen mit wenig oder gar keinem „normalen“ Text diese Auswertungskategorie nahezu wertlos ist.

Die einzig sinnvolle Alternative besteht bei der „Abfrage-Methode“ darin, jede Anzeige in klar erkennbare Teile zu gliedern, und die Beachtung für jeden Teil getrennt abzufragen.

Die „Anstreich-Methode“ liefert selbst die Grundlagen für die (zumindest konsistente) Kategorisierung im Institut.

Es ist meine Hoffnung, dass diese Ergebnisse und Überlegungen dazu beitragen, in Zukunft sogen. „Copy-Tests“ besser zu konzipieren.

Quellen:

- G+J Mafo-Daten-Dienst Nr. 92: "Marktforschungs-Methoden-Test", Interview und Analyse, Februar 1983, S. 90
- Koeppler, Karlfritz, u.a.: "Werbewirkungen definiert und gemessen", Heinrich Bauer Stiftung, Hamburg, 1974
- Lucas, Darrell B.: "The ABCs of ARF's PARM", Journal of Marketing, July 1960
- Schaefer, Wolfgang: „Copy-Tests und Copy Testing“, planung und analyse, Ausgabe 3/1992
- Schaefer, Wolfgang: "Recognition Reconsidered", marketing and research today, May 1995, ESOMAR, Amsterdam