Auf dem Weg zu erfolgversprechenden Parfüm-Produkttests - Ein Experiment

*Produkttests gehören zu den am weitesten verbreiteten und am häufigsten eingesetzten Untersuchungen im Bereich der Konsumgüter-Marktforschung. Der Autor ist seit nunmehr 20 Jahren Spezialist für derartige Tests und hat in dieser Zeit rund 2.000 Test für Waschmittel und Haushaltsreiniger, Nahrungs- und Genußmittel und Körperpflege-Mittel aller Art durchgeführt. Oft ging es bei diesen Tests auch um den Duft, wenn auch nur als ein (u.U. sehr wichtiger) Teilaspekt des Ganzen. Zunehmend wurden wir nun in letzter Zeit mit der einen oder anderen tastenden Frage konfrontiert, ob man nicht vielleicht auch **Parfums** testen könnte - also Produkte, bei denen der Duft die Produktleitung ausschließlich bestimmt. Diese vorsichtig formulierte Frage, aber auch das Drängen nach einer Lösung ist verständlich: Auf der einen Seite weiß man um die neuro-physiologischen Bedingungen, die das Riechen bestimmen und dass das Beschreiben von Düften schwierig bis unmöglich ist. Folgerichtig hatte man jahrzehntelang die allgemeine Vorstellung **„Düfte kann man nicht testen!"**. Auf der anderen Seite sieht der aufmerksame Beobachter, dass sich die größeren Konsumgüter-Hersteller, die sich in den letzten Jahren in der Duftbranche engagiert haben, ziemlich vorurteilslos und offensichtlich mit Erfolg an das Testen von Parfums gewagt haben.*

Wir waren schon seit längerem der Überzeugung, dass man Parfums testen kann, - wenn man es nur „richtig" macht und haben mit Hilfe der DRAGOCO ein Experiment durchführen können, um diese Vorstellung zu belegen. Was uns zu unserer Überzeugung gebracht hat, wie darauf fußend das Experiment angelegt wurde und was es an Ergebnissen und Erkenntnissen gebracht hat, soll im Folgenden geschildert werden.

Der falsche Ausgangspunkt: „Snifftests"

Die allgemeine Vorstellung, dass man „Düfte" nicht testen kann, rührt maßgeblich daher, dass man in der Vergangenheit mit wenig adäquaten Vorstellungen und Methoden an das Thema herangegangen ist. Der übliche „Snifftest", welcher im Bereich der Marktforschung weit verbreitet war und ist und mit Recht gescholten wurde, sah so aus, dass man einem Querschnitt von Konsumenten einen oder mehrere Düfte zum Riechen gab und sie fragte: „gefällt dies?" oder „welches gefällt besser?" und: „warum?".

Die Antwort konnte nur unbefriedigend sein, denn wir haben (zumindest in unserem Kulturkreis) keine Worte für Düfte, - außer vielleicht „angenehm", „frisch" und „zu intensiv". Diese wenigen Worte reichen naturgemäß nicht aus, um einen Duft zu beschreiben und schon gar nicht, ihn so zielgerichtet zu verbessern, dass er ein Erfolg wird.

Kein Wunder also, dass sich die Vorstellung, „Düfte kann man nicht testen", auf Seiten der meisten Parfumeure bis hin zu Vor-Urteilen verfestigt haben und entsprechende Versuche, Tests durchzuführen, kritisch gesehen, wenn nicht gar ganz abgelehnt werden.

Ist das die Lösung: „geschulte Nasen"?

Eine Antwort zur Schließung der „Sprachlücke" zwischen der Sprache der Verbraucher mit ihren wenigen Worten und dem Vokabular der Parfumeure lieferten jahre-, wenn nicht jahrzehntelang die Sensoriker - und diese Antwort sah aus, wie der Stein der Weisen: man schule Leute mit einer besonderen Riechbegabung, vorgegebene Düfte „exakt", d.h., mit vorgegebenen Worten nachvollziehbar und wiederholbar zu beschreiben.

Die vorgegebenen Worte werden in gemeinsamen Diskussionsrunden unter Zugrunde legen von Ankerreizen (also bestimmten, „reinen" und klar zu beschreibenden Düften) erarbeitet, die „geschulten Nasen" legen sich auf diese Worte fest und die Parfumeure lernen, sie zu übersetzen.

Was so faszinierend aussieht, hat allerdings einen Haken: wir Menschen können nun einmal nicht alle gleich gut riechen und eignen uns damit nicht alle gleichermaßen als Mitglieder eines solchen „Sensorik-Panels". Damit fußt eine solche „Beschreibung", und wenn sie noch so übereinstimmend von allen Mitgliedern des Panels vertreten wird, u.U. auf einem ziemlich

„schiefen“ Bevölkerungs-Querschnitt. Düfte werden aber durchaus an „Jedermann“ verkauft - auch an solche, die nicht besonders gut riechen können!

Natürlich wissen auch die Sensoriker, dass das „Beschreiben“ von Düften nur eine Seite der Medaille ist und dass die Kaufbereitschaft für ein bestimmtes Produkt und damit der nachhaltige Erfolg am Markt eher mit einer akzeptablen Bewertung und weniger mit einer exakten Beschreibung zu tun hat. Folgerichtig lassen sie ihre „geschulten Nasen“ dann auch nur „beschreiben“ und befragen parallel dazu eine Konsumentengruppen, wie die Düfte ganz generell gefallen, um ihre „Beschreibungen“ an marktrelevanten Akzeptanzmaßen verankern zu können.

Die Frage ist, was damit gewonnen wird: Das Marketing möchte ja nicht nur ein Produkt verkaufen, das „gefällt“ und daneben auf eine exakte Beschreibung durch ausgesuchte „Riech-Experten“ verweisen können, sondern auch vermitteln, dass sich mit diesen Düften eine ganze Reihe von Leistungsversprechen, Assoziationen, Stimmungen und Images verbinden.

Die Frage, ob dies gelingt, kann beantwortet werden. Nur, sie richtet sich eben nicht an das Sensorik-Panel, sondern an die Endverbraucher. Und damit schließt sich der Kreis:

- Gesucht wird ein Verbraucher-Test zur **Bewertung** von Düften, derart, dass man ablesen kann, ob bestimmte Marketing-Ziele und eine hohe Verbraucherakzeptanz erreicht werden kann;
- der gleichzeitig Stärken und Schwächen **beschreibt**, so dass die Parfumeure sie ggf. weiter optimieren und markfähig(er) machen können.

Unsere langjährige Erfahrung mit Produkttests zeigt, dass solcherlei Tests, sachgerecht angelegt und durchgeführt auch für Düfte und speziell auch für *Parfüms* erfolgreich eingesetzt werden können und Ergebnisse liefern, die beide Fragen beantworten können..

Die erste Voraussetzung: Experimental Design

Schon vor rund 30 Jahren, als wir begannen, für einen unserer ersten größeren Kunden, nämlich Colgate-Palmolive, *Geschirrspülmittel* zu testen, konnte sich keiner der für F&E zuständigen Chemiker vorstellen, dass die „normalen Hausfrauen“ Unterschiede zwischen zwei mit unterschiedlichen Konzentrationsgraden waschaktiver Substanzen angereicherten Geschirrspülmittel feststellen könnten. Konnten und können sie natürlich auch nicht, jedenfalls nicht jede einzelne für sich. Was war und ist in einem solchen Fall also zu tun? Wir konnten die Chemiker damals davon überzeugen, was der gelernte Physiker und seinerzeit berühmte deutsch-amerikanische Marktforscher *Alfred Politz* schon längst wusste:

1. man muss so etwas indirekt messen,
2. im Rahmen eines kontrollierten Versuchsplanes („experimental design“)
3. und auf der Basis einer ausreichend großen Stichprobe.

Was heißt das konkret?

ad 1) indirekte Messung

Sicherlich ist es unsinnig, Hausfrauen zwei verschiedene Geschirrspülmittel zu geben und zu fragen: „welches ist besser?“ und: „warum?“. Was sollen sie darauf antworten?

Indirekt messen heißt hier, dass man den Befragten eine Reihe von charakteristischen Eigenschaften und Leistungsmerkmale vorgibt und sie bittet, die Produkte anhand dieser Eigenschafschaften auf einer geeigneten Skala von „trifft überhaupt nicht zu“ bis „trifft voll und ganz zu“ einzustufen. Solche Statements reichen von

- hat eine angenehme Konsistenz
- schäumt gut
- riecht angenehm

über

- reinigt gründlich
- lässt das Geschirr wieder strahlen
- löst auch Fett besonders schnell
- greift die Hände nicht an

bis hin zu

- ist von hoher Qualität
- würde ich kaufen.

Konkret heißt dies, wir „messen", was die Testteilnehmerinnen an Qualitäten wahrnehmen oder wahrzunehmen glauben, nachdem sie es über einen ausreichend langen, alle Facetten der üblichen Verwendung berücksichtigenden Zeitraum hinweg ausgiebig getestet haben.

ad 2) Experimental Design

Um nun eine eindeutige Verbindungslinie zwischen den perzipierten Qualitäten auf der einen und den Produktformulierungen auf der anderen Seite herstellen zu können, muss man letztere systematisch variieren. Also: will man die unterschiedlichen Qualitäten verschiedener Konzentrationsgrade waschaktiver Substanzen messen, dann darf man in einer Versuchsserie nur diese verändern, und nicht auch bei Produkt A mit Konsistenz K_A, bei Produkt B mit Konsistenz K_B und bei dem dritten Produkt vielleicht auch mit einem anderen Duft arbeiten.

Gibt es mehrere solcher alternativer Rezeptur-Variationen, dann muss man einen (z.B. varianzanalytischen) Versuchsplan wählen, um mögliche Wechselwirkungen feststellen zu können. Aber es soll an dieser Stelle gar nicht so kompliziert gedacht werden. Wichtig ist festzustellen, dass man mit einer systematischen Variation des WAS-Anteils der Produkte so etwas herausbekommen kann wie:

- hoher Anteil WAS = schäumt gut = vermittelt sehr gut die Vorstellung von Wirkung = wenig mild
- mittlerer Anteil WAS = schäumt weniger gut = scheint auch ganz gut zu reinigen = etwas milder
- niedriger Anteil WAS = schäumt nicht = enttäuschende Reinigungswirkung = also harmlos=mild

ad 3) Große Stichproben

In seinem berühmten Aufsatz „The Illusion of an Image and the Excess Power of Groups"[1] hat *Alfred Politz* gezeigt, dass die statistische Gruppe mehr „weiß" als der einzelne, und wie sich aus der Vielzahl einzelner Eindrücke kumulativ ein Urteil über ein Produkt herauskristallisiert. In unserem Geschirrspülmittel-Fall sieht dies etwa wie folgt aus:

- Bei dem einen Produkt stellt die erste Befragte fest: „reinigt gut" und vergibt bei dem entsprechenden Item eine „6", um zu signalisieren „trifft weitgehend/fast vollständig zu"; die nächste urteilt ähnlich, was wieder zu einer „6" führt und so geht es weiter mit Urteilen, wie etwa 7,6,5,5,6,6,7,5,4,5,6,... was sich dann insgesamt vielleicht zu einem Mittelwert von sagen wir mal 5.6 addiert.

- Betrachtet nun das nächste Produkt und unterstellt man vielleicht auch, dass dieses von denselben Personen getestet worden wäre[2], wäre zum Beispiel herausgekommen, dass die 1., 5., 6., 8., 12.,13. und weitere... Personen gar keine Unterschiede hinsichtlich der Reinigungswirkung festgestellt haben. Aber die anderen mögen das ungute Gefühl gehabt haben: „so toll ist das nicht" und mögen demnach einen oder zwei Punkte niedriger auf der 7er-Skala angekreuzt haben. Das hätte dann genau dazu

[1] abgedruckt in Hugh S. Hardy: „The Politz Papers - Science and Truth in Marketing Research", AMA 1990 und von ihm bereits 1963 auf dem Amsterdamer ESOMAR-Kongress vorgetragen

[2] aus Gründen, die wir auf dem 207.ESOMAR-Seminar 1995 vorgetragen haben, bevorzugen wir gegenüber dem obigen Gedankenexperiment allerdings normalerweise **monadische** Tests, bei denen jede Testteilnehmerin nur **ein** Produkt zu testen hat.

geführt, dass zwar viele Einzelne keinen Unterschied festgestellt haben - insgesamt aber doch ein Unterschied herauskommt, denn der Mittelwert von (angenommener Weise) 5.2 signalisiert: dies Produkt schneidet bezüglich des Items „reinigt gründlich“ signifikant schlechter ab!

Wichtig ist, dass die Testgruppen groß genug sind, so dass man in allen Gruppen einen ausreichend großen und gleichmäßigen Anteil mit „regelmäßigen“ und mit „gelegentlichen“, Markenverwender der „schonend reinigenden/umweltfreundlichen“, „milden“ wie der „kräftigen“ Marken, aus Haushalten mit und ohne Geschirrspüler und dergleichen mehr hat, was die Verwendungsgewohnheiten und Ansprüche an Geschirrspülmittel bestimmen mag. Das Stichwort hierfür ist: die Strukturgleichheit der Stichproben ist sicherzustellen *(matched samples)*.

Wichtig ist auch, dass die Testzeit lang genug währt, so dass sich Urteile aufbauen und verfestigen können (der erste Eindruck ist oft ein anderer als der nach einiger Zeit und Gewohnheit). Schließlich muss man eine ausreichend große Produktmenge zur Verfügung haben, um auch ausgiebig testen zu können.

Interessant ist, dass man nicht sicher sagen kann, dass die einzelne Testteilnehmerin in jedem Falle zu einem späteren Zeitpunkt genauso urteilen würde, wie beim ersten Mal und dass sich trotzdem - *ceteris paribus* - die gleichen Testergebnisse wieder einstellen würden, würde man den Test zu einem solchen späteren Zeitpunkt zu den gleichen Bedingungen wiederholen. Wundersam? Nein: „The Excess Power of Groups“ !

Obwohl dieses Beispiel erst einmal nichts mit dem Thema „Dufttests“ zu tun hatte, habe ich es mit Absicht etwas ausgewalzt, um daran das grundlegende Prinzip darzulegen. Denn heute dürfte es nur wenige geben, die das eingangs genannte Vorurteil noch hegen und bezweifeln, dass man Geschirrspülmittel testen kann: natürlich geht so etwas; es ist doch seit langem bewährte Praxis...

Die zweite Voraussetzung: geeignete Beurteilungskriterien

Es gibt noch einen zweiten Bereich, in dem nicht von vorn herein „klar“ war, dass man hier auf erfolgversprechende Weise Produkttests durchführen könnte: seit Anfang der 80iger Jahre führen wir Jahr für Jahr Testserien mit Produkten durch, die in besonderem Maße von Geschmack und Aroma bestimmt werden - Zigaretten und Kaffee.

Das wesentliche Problem war dabei nicht so sehr die Frage der Testanlage. Hier war uns klar, dass man wiederum mit unserem bewährten Ansatz Erfolg haben würde. Das Problem war vielmehr das Finden geeigneter Beurteilungskriterien, also solcher Statements, die geeignet waren, die besonderen Produkteigenschaften und Leistungen so zu beschreiben, dass sich Verbraucher etwas darunter vorstellen konnten - und die auch noch einen Bezug zur von sensorischen Begriffen geprägten Welt der Produktentwickler haben.

Zunächst haben wir uns *per judgement* auf Eigenschaften gestützt, die in Diskussionsrunden zwischen Marktforschung, Marketing und Produktentwicklung formuliert und festgelegt worden waren. Dann haben wir ein Experiment durchgeführt, bei dem eine ganze Reihe verschiedener Kaffee-Varianten mit Hilfe von mehr als 90 denkbarer oder „am Rande“ liegender Statements zu bewerten waren. Faktoren- und Regressions- und Varianzanalysen halfen, diejenigen Eigenschaften einzugrenzen, die man als *Key Items* bezeichnen könnte: das sind solche Items,

- die gut zwischen verschiedenen Kaffee-Sorten zu unterscheiden vermögen;
- und einen relativ hohen Einfluss auf die Erklärung von Geschmackspräferenzen, Qualitätsvorstellungen und letztendlich „Kaufbereitschaft“ haben.

Die meisten dieser Items beschreiben dabei Produkteigenschaften, über deren Charakteristik sich Verbraucher wie Experten gleichermaßen einig sein können. Was „bitter“ ist, das weiß oder glaubt man auf beiden Seiten gleich gut zu verstehen; auch über andere Dinge gibt es oft keine weit auseinandergehenden Meinungen. Manchmal stellen sich die Experten aber doch die Frage, was eine bestimmte Aussagen zu bedeuten hat, etwa „vollmundig“ oder „milder Geschmack“: was ist das genau?

Hier half uns einmal mehr Politz'sches Gedankengut, den einen oder anderen Experten auf Kundenseite davon zu überzeugen, dass man einerseits manchmal „um die Ecke“ denken, andererseits die sensorische Seite verlassen und über systematische Versuchsreihen die „Sprache der Verbraucher“ kennenlernen muss, wenn es nicht - wie z.B. beim Thema „Wein“ umgekehrt gelingt, den Verbrauchern die Expertensprache nahezubringen.

So hat *Politz* z.B. in seinem Interview mit Eva Bartos, abgedruckt in den *Marketing Masters*[3], anschaulich geschildert, wie er einst herausgefunden hatte, dass das Statement, welches am besten die Präferenz für eine Zigarette charakterisierte, das Statement *„it has more tobacco taste“* war. Nur, Zigaretten haben alles, nur keinen *tobacco taste*; gleichwohl wurde dieses Statement im Folgenden verwendet, um in systematischen Testreihen festzustellen, was man objektiv in der Behandlung des Zigaretten-Tabaks tun müsste, um dem Verbraucher-Ideal: *„has more tobacco taste“* möglichst nahe zu kommen.

Mit anderen Worten, es ist wichtig, einen Übersetzungs-Schlüssel für die Worte der Verbraucher zu finden - und das funktioniert am besten über Versuchsreihen, bei denen man Produkte systematisch und in bekannter Weise variiert und dann genau betrachtet, wie diese Varianten von den Verbrauchern bewertet werden. Durch Ziehen von Verbindungslinien zwischen dem Produkt-*input* zu den mit Hilfe geeigneter Statements gewonnen Beschreibungen und Bewertungen als *output* gelangt man genau zu dem Übersetzungscode, der es dann gestattet, Produkte zielgerichtet (weiter) zu entwickeln.

Dass man für solcherlei Übersetzungsarbeit nicht unbedingt ein sensorisch geschultes Panel dazwischen schalten muss, hat uns ein Kollege aus der Zigaretten-Industrie einmal sehr anschaulich erklärt:

- Meine Produktentwickler sprechen und verständigen sich in einer Spezialsprache - sagen wir einmal: *französisch*;
- die Verbraucher verstehen aber nur *deutsch*;
- was nützt es mir, jetzt ein *englisch* sprechendes Panel dazwischen zu schalten um die Verbrauchersprache anzureichern, wenn ich auch das wieder - in langen Testreihen - übersetzen muss?

⇒ Dann kann ich doch gleich lange Testreihen mit *deutschen* Begriffen durchführen und die Experten bitten, ihre *französischen* Vokabeln und das Wissen um die Rezepturen und Formeln dagegen zu setzen - schon habe ich die Übersetzung!

Der Einstieg in das Duftthema

Die meisten der von uns zu testenden Konsumgüter haben einen charakteristischen *Duft*, der helfen soll, die wesentlichen Produkt-Benefits zu unterstreichen und die verschiedenen Marken

[3]The Marketing Masters: Interviews mit den Founding Fathers der Markt- und Sozialforschung, 1936, wieder aufgelegt durch die AMA 1991

zu differenzieren, seien es nun Allzweckreiniger & Geschirrspülmittel, Waschmittel & Weichspüler, Badezusätze oder Duschbäder, Haar- oder Körperpflegemittel. Zur Bewertung dieser Düfte werden üblicherweise folgende Fragen gestellt und untersucht

- ob der Duft gefällt
- ob die Intensität gut ausbalanciert ist, der Duft ausreichend lange oder gar zu lange anhält
- und ob er im richtigen Verhältnis zu zwei charakteristischen Polen eingestuft wird: „unaufdringlich/dezent“ und „frisch“ (wobei diese beiden Pole in bestimmten, produkttypischen Facetten zu beschreiben sind).
- Und schließlich wird untersucht, ob der Duft in der Lage ist, die wesentlichen Produktleistungsversprechen zu unterstützen oder zu unterstreichen[4].

Ist z.B. ein neuer Duft für ein Anti-Schuppen-Shampoo einer Marke wie Timotei oder NIVEA zu entwickeln, so soll dieser ja in erster Linie die Vorstellung dieses Produkt-Konzeptes unterstützen:

- mild sein
- aber auch überzeugend wirksam.

Der Parfümeur macht sich nun daran und entwickelt eine Vielzahl denkbarer Lösungen - vielleicht auch solche, die in der einen oder anderen Richtung (graduell) über das Ziel hinausschießen. Tests, welche die obigen Bewertungskriterien verwenden, würden dann zeigen, welche der Düfte die gewünschten Benefits und die Eignung für die Marke, unter denen das Produkt später verkauft werden soll, am ehesten unterstreichen (natürlich muss der Duft auch gefallen und Kaufbereitschaft auslösen helfen). Man macht also letztendlich nichts anderes als Alfred Politz mit seinen Zigaretten: Man testet solange, bis man die Variante gefunden hat, die - im übertragenen Sinne - den besten *tobacco taste* hat!

Allerdings sind wir auch hier in den 30 Jahren seit Politz inzwischen ein Stückchen weiter gekommen, denn ebenso, wie beim Testen von Kaffee oder Zigaretten wird man zusätzlich noch mit ausgewählten und ausgetesteten Items arbeiten, die eine möglichst differenzierte Duftcharakteristik zu liefern in der Lage sind. Dies beginnt bei der Frage der Intensität und anhaltenden Wirkung, berücksichtigt das „richtige“ Verhältnis von Unaufdringlichkeit und doch markantem und unverwechselbaren Eindruck (meist mit „Frische“ assoziierbar) und stützt sich darüber hinaus auf weitere geeignete und den Eindruck abrundende „Deskriptoren“.

Der letzte Punkt vermischt sich mit der Frage nach den „Benefits“, wenn der Duft selbst zum Haupt-Benefit wird, etwa bei Deos, - seien es nun Mild-Deos, Wirk-Deos/Antitranspirantien (bei denen man noch einen gewissen Zusatznutzen unterstellt), oder ganz besonders bei Parfüm-Deos oder gar EdTs und Parfüms (wo es letztendlich nur noch um den Duft geht). Und so gilt es, kreativ zu werden und möglichst viele Facetten bei der Wahl geeigneter Bewertungskriterien einzubeziehen.

In den 80er-Jahren wurde einmal ein Verfahren bekannt, das in England entwickelt wurde und sich „Fragrance Mapping“ nannte[5]. Bei diesem Verfahren wurden manchmal über hundert „Eigenschaften“ verwendet, um alle existierenden Düfte in einem bestimmten Produktfeld zu

[4]siehe hierzu auch Ivor Shalofskys Referat auf dem ESOMAR-Seminar on Flavours & Fragrances, 1993 in Köln
[5]Der Autor dieses Verfahrens hieß m.E. Allan Frost; leider ist mir keine Original-Quelle bekannt, die ich zitieren könnte, um die Idee dieses ideenreichen Forschers hier adäquat zu würdigen

charakterisieren; sie reichten von beschreibenden Kriterien über Zielgruppen-Beschreibungen, Produkt-Assoziationen, Jahreszeiten, Naturbeschreibungen, Farben, Personenbezügen, Stimmungen etc., pp. Das ist eine „Fundgrube“ für Itembatterien:

Eigenschaften

frisch	pflegend	modern	männlich	markant	intensiv
langanhaltend	aufdringlich	würzig	herb	alltäglich	weiblich
billig	positiv	klassisch	sinnlich	kühl	aktiv
aggressiv	leicht	jung	natürlich	elegant	fruchtig
chemisch	exotisch	faszinierend	leicht	anregend	unkonventionell
fremdartig	dezent	interessant	mild	blumig	süßlich
langweilig	anders	wertvoll	erotisch	dynamisch	schwer
beruhigend	Phantasie anregend...				

Zielgruppe

für junge	alte	Babys	Hausfrauen...

Produkte

passt zu....	Shampoo	Seife	EdT...

Jahreszeiten

Frühling	Sommer	Herbst	Winter

Farben

rot	blau	grün	gold	orange

Natur

Meer	frische Brise	Sommertag	Blumenwiese	Wald	Abenteuer
Herbstspaziergang ...					

bekannte Personen

Robert Redford	Caroline v. Monacco	Steffi Graf...

Ein Experiment mit Eau de Toilettes

Im Rahmen einer Neuanwerbungsaktion für unser Produkttest-Panel haben wir 1996 einmal eine Reihe von EdTs getestet. Üblicherweise führen wir mit neu angeworbenen Haushalten erst einmal einen Probetest durch, bei dem die Teilnehmer das Testverfahren kennenlernen und wir uns davon überzeugen können, dass sie auch eines Tages einmal antworten werden, wenn wir sie in einen „ernsthaften“ Test einbeziehen.

DRAGOCO hatte uns dazu freundlicherweise 13 verschiedene EdTs hergestellt und in kleinen Probierfläschchen abgefüllt. Somit konnten wir den neuen Panelmitgliedern gleich etwas Interessantes bieten - und im Gegenzug hofften wir, demonstrieren zu können, ob und in welchem Maße „klassische“ Produkttests geeignet sein könnten, solcherlei Düfte adäquat zu bewerten.

Bei den 13 Düften handelte es sich einmal um „Tresor“ als *benchmark* und um je 3 Düfte, die vier DRAGOCO-Parfumeure eigens für diesen Test kreiert hatten.

Die Stichprobe für dieses Test-Experiment war, anders als bei üblichen Tests, in diesem Falle nicht auf die Zielgruppe der EdT-Verwender oder gar der Verwender von oder Interessenten an bestimmten Duftrichtungen ausgerichtet: es wurde der Querschnitt der Frauen angesprochen, so wie er für das Panel angeworben worden war (mit einem Schwerpunkt bei jüngeren Frauen bis max. 34 Jahren). Gleichwohl versprachen wir uns von diesem Experiment einen Aufschluss darüber, ob ein „klassischer“ Test generell funktionieren kann oder nicht.

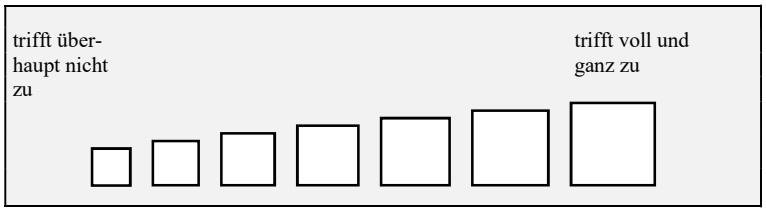
Der Fragebogen enthielt offene Fragen zu „Likes“ und „Dislikes“ (die allerdings nicht ausgewertete wurden), die Frage nach der Intensität des Duftes und eine Itembatterie, die folgende Statements umfasste:

- natürlich, frisch, feminin, kühl, blumig, süßlich, holzig, intensiv, markant, spritzig, herb, erotisch, krautig, dezent, avantgardistisch, langanhaltend, aufdringlich, leicht, unkonventionell, interessant, angenehm, modern, exotisch, langweilig, würzig, kraftvoll, gewöhnungsbedürftig, für besondere Anlässe, in die heutige Zeit passend, elegant *und* wie aus 1001 Nacht.

Last but not least wurde das Gefallen ermittelt und zwar mit den Statements

- der Duft ist angenehm
- der Duft passt ... überhaupt nicht zu mir, ..., voll und ganz zu mir
- diesen Duft würde ich spontan bestimmt nicht kaufen, ..., bestimmt kaufen

Alle Statements waren anhand einer 7er-Skala des folgenden Typs einzustufen:



Die Auswertung des Tests geschah in tabellarischer Form, derart, dass wir für die obigen Statements Mittelwerte berechneten und auswiesen. Nun ist es so, dass solcherlei „längliche“ Itembatterien naturgemäß eine gewisse Redundanz aufweisen; einige Items bedeuten ähnliches oder werden von den Befragten in ähnlicher Weise verstanden. Das ist z.T. gewollt, will man doch die verschiedenen Facetten eines Begriffes wie „frisch“ möglichst differenziert erfassen - etwa, um herauszufinden, ob es eher in die Richtung „frische Brise“ oder „Blumenfrische“ geht.

Üblicherweise führen wir deshalb eine *Faktoren-Analyse* durch, die den inneren Zusammenhang in einer derartigen Itembatterie aufdeckt und Gruppen von Items ausweist, die in besonderer Weise miteinander korrelieren. Hier erhielt man ganz konkret folgende Faktoren[6]:

	Ladung
Faktor 1:	
- in die heutige Zeit passend	0.73
- modern	0.69
- interessant	0.67
- erotisch	0.51
- aufdringlich	-0.58
- gewöhnungsbedürftig	-0.63
- langweilig	-0.81
Faktor 2:	
- dezent	0.82
- leicht	0.78
- natürlich	0.57
Faktor 3:	
- holzig	0.80
- krautig	0.76
- würzig	0.70
- herb	0.54

	Ladung
Faktor 4:	
- kühl	0.83
- frisch	0.70
- spritzig	0.65
Faktor 5:	
- für besondere Anlässe	0.81
- elegant	0.76
Faktor 6:	
- langanhaltend	0.90
- intensiv	0.75
- kraftvoll	0.48
Faktor 7:	
- blumig	0.78
- süßlich	0.70
- feminin	0.55

	Ladung
Faktor 8:	
- exotisch	0.79
- wie aus 1001 Nacht	0.75
Faktor 9:	
- unkonventionell	0.76
- avantgardistisch	0.74
Faktor 10:	
- markant	0.77

[6]Die Faktoren-Lösung basiert immer auf den Daten, die Basis für die Berechnung waren. Hier waren es die Bewertungen für die 13 getesteten Düfte. Würde man alle wichtigen Parfüms auf gleiche Weise mit Hilfe derselben Items bewerten lassen, so erhielte man eine Struktur, die für EdTs/Parfums in allgemeiner Weise gelten würde.

Es ist denkbar und des Öfteren üblich, dass man den einzelnen Faktoren „Namen“ gibt, etwa indem man ihnen entsprechend der Bedeutung der unter ihnen subsumierten Items eine prägnante Überschrift zuweist. Wir verzichteten hier ganz bewusst darauf, weil wir nicht wollten, dass sich diese Namen in diesem frühzeitigen Experimentier-Stadium verselbständigen.

Man spricht dann nämlich ganz schnell von „*Der* Frische“ oder „*Der* Exotik“ und jedermann glaubt zu wissen, worum es geht. Aber hier war es erst das Ziel, das Verfahren ganz generell einem Praxis-Check zu unterziehen und dann, geeignete Key Items zu finden; noch war ja nicht klar, ob dies alles funktionieren könnte.

Ein erster Check, der zeigen sollte, ob wir eine Item-Auswahl getroffen hatten, die geeignet war, die verschiedenen EdTs adäquat zu bewerten, lag darin, zu überprüfen, ob und inwieweit hiermit die Vorstellungen

- das ist ein angenehmer Duft
- dieser Duft passt zu mir
- *und*: ein EdT mit diesem Duft würde ich spontan kaufen

erklärt werden können.

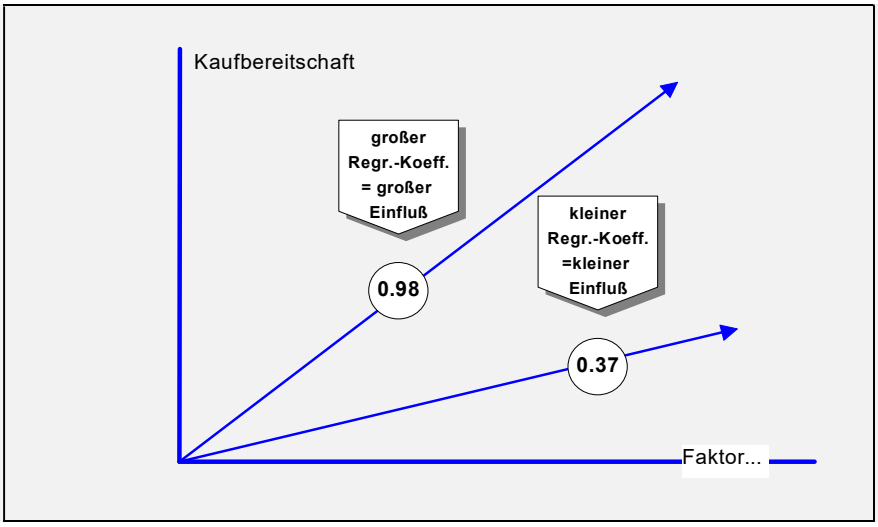
Das Verfahren, welches dieses überprüft, ist die Multiple Regressionsanalyse (MRA). Sie liefert zunächst einmal Regressionskoeffizienten, der den Zusammenhang zwischen den Faktoren (als erklärende Variable) und den zu erklärenden Messwerten anzeigt. Hieraus lässt sich das sog. *Bestimmtheitsmaß* ableiten, das aussagt, wie viel der Varianz in den allgemeinen, übergeordneten Bewertungen (denn jede Befragte hat eine andere Vorstellung davon, ob es sich z.B. um einen „angenehmen Duft“ handelt..) erklärt oder *bestimmt* wird. Diese Bestimmtheitsmaße sahen wie folgt aus:

⇒ das ist ein angenehmer Duft	**75.5**%
⇒ dieser Duft passt zu mir	**73.7**%
⇒ ein EdT mit diesem Duft würde ich kaufen	**71.1**%

Das Ergebnis zeigt, dass wir auf einem guten Weg waren. Sagt es doch, dass die Vorstellung, dass eines der EdTs ein „angenehmer Duft“ war, zu drei Vierteln durch die Auswahl unserer vorgegebenen Deskriptoren und Assoziationen beschreibbar war. Selbst die „Kaufbereitschaft“ war zu gut 70% hiervon bestimmt; nur ein „Rest“ von rd. 30% war dann noch abhängig von nicht erfassten Kriterien, wie etwa von „fehlenden“ Duftbewertungskriterien, und sicherlich ganz nachhaltig, von Marken-Namen/Image und Preis. Alles dies gilt auch für andere Konsumgüter-Produktfelder und so erreichen wir auch in anderen Produkttests selten Werte, die hierüber liegen.

Aufgrund dieser Werte waren wir also guten Mutes bei DRAGOCO’s Parfumeuren angetreten, um die Bewertungen der von ihnen kreierten EdTs vorzustellen.

Bei der Darstellung wählten wir eine Gruppierung der Düfte nach deren relativen Ähnlichkeit hinsichtlich der Faktoren, die für ein Gesamturteil am wichtigsten gewesen waren. Denn die MRA liefert neben dem oben angesprochenen Korrelationskoeffizienten auch noch sog. Regressionskoeffizienten, welche die relative Wichtigkeit der einzelnen Faktoren beziffern. Diese Regressionskoeffizienten zeigen an, wie „Gefallen“ ansteigt, wenn ein Duft bezüglich eines bestimmten Faktors verbessert werden kann. Dargestellt sei dies anhand folgender Grafik:



Dabei kann durchaus der Eindruck „dies ist ein angenehmer Duft“ von anderen Dingen bestimmt werden als die Vorstellung „das passt zu mir“ oder „würde ich kaufen“, zumindest können sich die Gewichte verschieben (so etwa, wie die Vorstellung: „Mercedes ist ein qualitativ hochwertiges Auto“ von anderen Dingen abhängen kann als die Vorstellung, einen solchen Wagen kaufen zu wollen).

Dementsprechend lieferten die getrennt mit Bezug auf die drei allgemeinen Akzeptanzmaße bezogenen MRA auch leicht unterschiedliche Ergebnisse:

		Regressionskoeffizienten in Bezug auf...		
		angenehm	**passt zu mir**	**Kaufbereitschaft**
Faktor 1:	modern/interessant	**0.71**	**0.75**	**0.76**
Faktor 2:	dezent/natürlich	0.44	0.41	0.40
Faktor 3:	holzig/krautig	-0.04	-0.10	-0.13
Faktor 4:	kühl/frisch	0.17	0.20	0.22
Faktor 5:	etwas Besonderes	0.23	0.24	0.26
Faktor 6:	anhaltend/intensiv	-0.03	0.00	0.06
Faktor 7:	blumig/feminin	0.03	-0.01	-0.02
Faktor 8:	exotisch	0.10	0.13	0.17
Faktor 9:	unkonventionell	0.08	0.03	0.03
Faktor 10:	markant	-0.03	0.01	0.02

Der „wichtigste“ Faktor war demnach der erste Faktor, sicherlich auch deshalb, weil er sich auf Items stützt, die mehrheitlich einen deutlich hedonistischen Charakter haben. An zweiter Stelle folgte Faktor 2, dann 5 und 4, etwas danach Faktor 8 - und in kontraproduktiver Weise Faktor 3. Praktisch keine Bedeutung für das Zustandekommen eines Gesamtgefallens hatten die Faktoren 6, 7, 9 und 10.

Allerdings sehen wir aufgrund der Ergebnisse einer anderen Analyse, nämlich einer Varianz-Analyse, dass selbst einige dieser zuletzt genannten Faktoren durchaus in der Lage waren, deutlich zwischen den getesteten Düften zu differenzieren. Hier die ermittelten F-Werte und die Spannweite in den erzielten Ergebnissen:

		F-Wert	Spannweite
Faktor 1:	modern/interessant	2.87	1.52
Faktor 2:	dezent/natürlich	1.61	1.15
Faktor 3:	holzig/krautig	3.13	1.16
Faktor 4:	kühl/frisch	4.37	1.61
Faktor 5:	etwas Besonderes	1.98	1.25
Faktor 6:	anhaltend/intensiv	1.66	1.01
Faktor 7:	blumig/feminin	4.62	1.55
Faktor 8:	exotisch	3.66	1.55
Faktor 9:	unkonventionell	1.38	0.82
Faktor 10:	markant	1.08	0.76

Bis auf die letzten beiden, sind alle F-Werte auf mindestens 90%-Niveau signifikant, die Werte über 2.0 gar auf 99%-Niveau. Das heißt also, auch die Faktoren 6,7 und 8 hatten eine gute Berechtigung und sollten nicht aufgegeben werden, bevor weitere Messreihen mit anderen Düften vorliegen.
Selbstverständlich lassen sich diese regressions- und varianzanalytischen Betrachtungen auf Item-Ebene fortsetzen, um ideale Repräsentanten für die Faktoren zu finden, weniger geeignete Items zu eliminieren und um ggf. neue und bessere aufzunehmen. Dies wir umso besser gelingen, als in zukünftigen Testserien noch weitere, den Duftmarkt bestimmende Marken und Duftrichtungen mit aufgenommen worden sein werden. Bekanntlich war die hier vorliegende Testreihe neben dem bekannten *Tresor* erst einmal nur neue Kreationen enthalten (auch wenn diese sich vermutlich an Bekanntes angelehnt hatten). Je repräsentativer eine solche Datenbasis letztendlich ist, desto allgemeingültiger werden die aus ihr ableitbaren Erkenntnisse.

Kommen wir nun zu den Ergebnis-Darstellungen und den Ergebnissen selbst. Grundsätzlich wäre es denkbar gewesen, die Düfte aus mehreren „Blickwinkeln“ zu betrachten:

- „eindimensional“
  - welche werden als blumig beschrieben,
  - welche als frisch?
- „mehrdimensional“
  - welche werden als frisch **und** blumig beschrieben,
  - welche als frisch **und** männlich?
- „komplex“
  - welche ähneln sich bezüglich aller „wichtigen“ Kriterien,
  - oder bezüglich aller „differenzierenden“,
  - oder gar „über alles“?

Diese vergleichenden Darstellungen bekommt man, indem man die Düfte bezüglich bestimmter Kriterien „clustert[7]“. Hier die verschiedenen Ergebnisse, die herauskommen, wenn man unterschiedliche Kriterien als Basis für die Gruppierung nimmt:

[7]mit Hilfe einer sog. Cluster-Analyse lassen sich auf statistischem Wege Gruppen von Objekten bilden, die sich hinsichtlich vorgegebener Eigenschaften möglichst ähnlich sind,- und bei denen sich die einzelnen Gruppen genau hinsichtlich dieser Eigenschaften möglichst deutlich unterscheiden

Clusterung nach:

modern/interessant	Faktor 7	Faktoren 1,3,4,7 und 8	alle Faktoren
Tresor	Tresor	Tresor	Tresor
Dunette	Pastello	Pastello	Marylin
Vital	Marylin	Marylin	Pamela
Anita	Anita	Pamela	Cannes
Plaisir	Pamela	Anita	Geena
Cannes	Blue Hat	Geena	Vital
Geena	Geena	Vital	Plaisir
Marylin	Franziska	Cannes	Franziska
Pamela	Plaisir	Plaisir	Dunette
Pastello	Vital	Dunette	Pastello
Vanilla Flower	Cannes	Fransizka	Anita
Blue Hat	Vanilla Flower	Vanilla Flower	Blue Hat
Franziska	Dunette	Blue Hat	Vanilla Flower

Man erkennt an dieser Darstellung, dass sich die Düfte durchaus unterschiedlich gruppieren, wenn man unterschiedliche Gruppierungsmerkmale verwendet. Dies ist im Prinzip die Frost'sche Idee des „Fragrance-Mappings":

- Je nach der Vorstellung, die mit Hilfe eines Duftes zu evozieren ist, clustert man die am Markt vorhandenen Düfte,
- lernt daraus, welche Lösungen einer bestimmten Vorstellung am nächsten kommen, welche weit weg führen (und alle Zwischentöne dazwischen)
- und kann dann zielgerichtet auf eigene, neue Lösungen zusteuern.

Das heißt, hat man erst einmal einen Satz an Düften zur Verfügung, und diese anhand der verschiedensten Deskriptoren und Assoziationen auf beschriebene Weise bewerten lassen, dann lässt sich diese „doppelte Vielfalt" verwenden, um in unterschiedlichster Weise Verbindungslinien zwischen dem *input*, nämlich den Parfumrezepturen und dem *output*, nämlich den perzipierten Verbrauchervorstellungen ziehen; - ein vielfältig nutzbares Übersetzungsprogramm!

Nachfolgend seien die Faktor-Mittelwerte für die 13 EdTs dargestellt; die Düfte sind dabei in der Reihenfolge gelistet, wie sie sich aus der Clusterung nach den „wichtigsten" Faktoren ergibt:

	Faktor 1 modern/ interessant	Faktor 2 dezent/ natürlich	Faktor 3 holzig/ krautig	Faktor 4 kühl/frisch	Faktor 5 etwas Besonderes	Faktor 6 anhaltend/ intensiv	Faktor 7 blumig/ feminin	Faktor 8 exotisch	Faktor 9 unkonven-tionell	Faktor 10 markant
Tresor	4.7	3.8	1.9	3.6	4.3	4.2	5.2	3.3	3.2	3.8
Pastello	4.1	3.3	1.8	3.2	3.5	4.4	4.9	3.1	3.0	3.5
Marylin	4.7	3.6	2.2	3.4	3.9	4.8	5.0	3.3	3.3	4.0
Pamela	4.6	3.5	2.1	3.5	3.7	4.5	4.8	3.2	3.0	4.0
Anita	4.6	3.6	2.3	3.5	3.5	4.3	4.9	3.4	3.3	3.8
Geena	4.7	3.6	2.1	3.6	3.8	4.2	4.4	2.7	3.0	3.5
Vital	4.6	4.1	2.1	3.8	3.8	4.2	4.5	2.9	3.3	4.1

Cannes	4.3	3.2	2.3	3.1	3.6	4.6	4.6	2.8	2.9	3.9
Plaisir	4.3	3.6	2.5	3.9	3.2	4.2	4.5	2.6	3.1	3.5
Dunette	4.6	3.6	2.7	3.8	3.9	4.7	4.3	3.2	3.3	4.2
Franziska	3.9	3.7	2.6	3.5	3.5	4.2	4.4	2.7	3.0	3.9
V. Flower	3.9	2.8	2.9	2.9	3.3	4.9	4.2	3.4	3.4	4.1
Blue Hat	4.1	3.2	2.2	2.9	3.7	4.6	5.3	3.9	3.0	4.1

Dass Ähnlichkeit nicht immer bedeutet, dass die Akzeptanz gleich hoch ist, zeigen die von den einzelnen Varianten erzielten Werte bezüglich der allgemeinen Akzeptanzwerte „ist angenehm“, „passt zu mir“ und „würde ich kaufen“:

	angenehm	passt zu mir	Kaufbereitschaft
Tresor	**5.1**	4.6	4.2
Pastello	4.1	3.7	3.3
Marylin	**4.9**	4.3	4.0
Pamela	4.6	4.1	3.9
Anita	4.6	4.0	3.7
Geena	**4.8**	4.3	4.1
Vital	4.7	4.2	3.9
Cannes	4.4	3.7	3.2
Plaisir	4.2	3.7	3.4
Dunette	**4.7**	4.0	3.7
Franziska	4.0	3.4	3.1
V. Flower	3.8	3.4	2.9
Blue Hat	**4.0**	3.6	3.2

Selbstverständlich liefert der Test nun noch weitere Möglichkeiten der Diagnostik und damit des „fine tunings“ der verschiedenen Düfte. Einmal ist die Darstellung hier auf die Faktor-Mittelwerte verkürzt worden; selbstverständlich kann man nun daran gehen und die genauen Unterschiede, besondere Stärken und/oder mögliche Schwächen anhand der einzelnen Items betrachten: ist ein Duft etwa zu süßlich, etwas zu holzig oder würzig, cool oder intensiv?

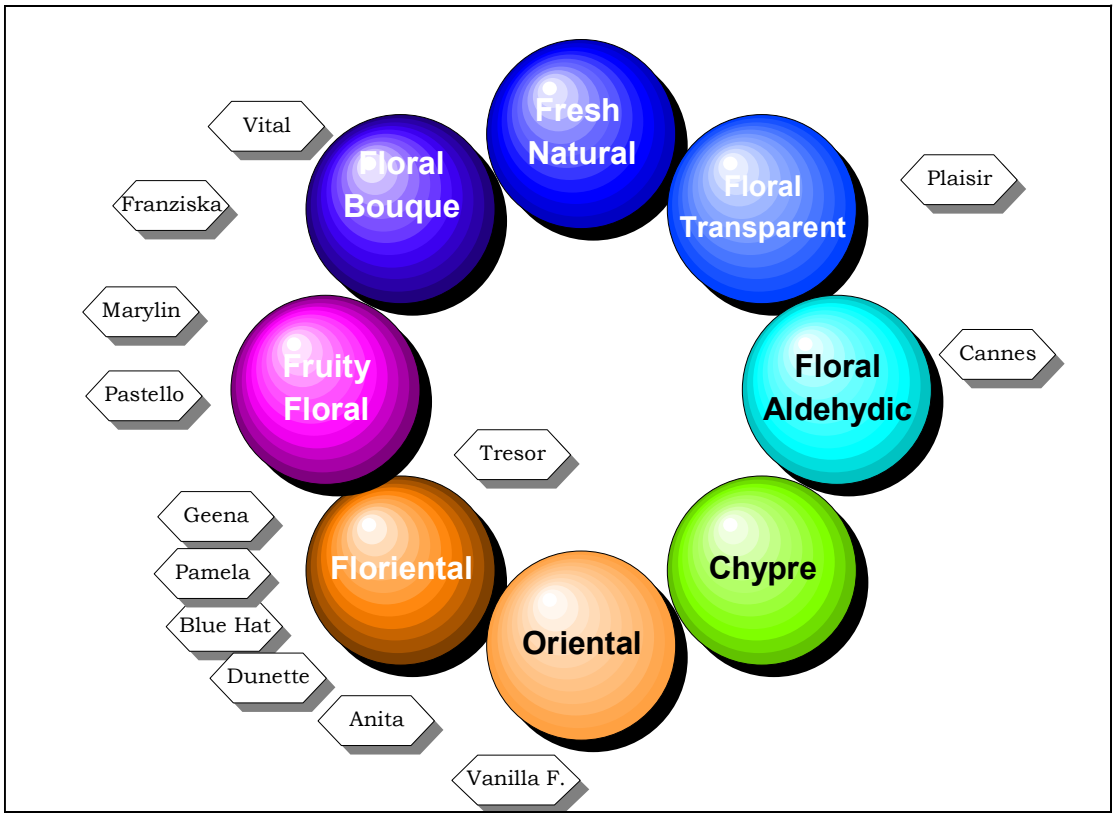
Gerade die Intensität wurde noch zusätzlich untersucht: ist der Duft im Einzelnen etwas oder viel zu stark, gerade richtig oder zu lange anhaltend? Und wie entwickelt sich der Eindruck des Gefallens vom ersten Eindruck bis hin zu längerer Anwendung? Gerade Letzteres ist immer wieder ein interessanter Punkt: Gibt es zwischen erstem und dem nachhaltigen Eindruck einen größeren Einbruch in der Bewertung, muss man sich Gedanken machen zur Balance von Kopf- und Herznote.

Fazit und Ausblick

Als Summary unseres Experimentes, das auf langjährigen Erfahrungen mit dem Testen funktionaler Produkte, und der Bewertung und Beschreibung von deren wesentlichen Charakteristika und Benefits ebenso basiert wie auf Geschmacks- und Aroma-Tests, Tests mit „bedufteten“ Produkten und reinen „Duft-Snifftests“ beruht, möchten wir darstellen, wie sich die 12 neu kreierten Düfte nach Aussage der involvierten Parfumeure im DRAGOCO-Octagon platzierten.

Der interessierte Leser mag aus diesem *input* und unseren Ergebnissen als *output* ablesen, ob sich mit Hilfe solchermaßen gewonnener Messwerte die mehrfach angesprochenen Verbindungslinien ziehen lassen und ein Übersetzungscode geschaffen werden konnte zwischen der Sprache und den Vorstellungen der Parfumeure und den Assoziationen und schließlich der Akzeptanz & Kaufneigung auf Seiten der Verbraucher.

Wir sind der Meinung: das Experiment weist den richtigen Weg. Verbesserungen mit Hilfe weiterer, die Assoziationsmöglichkeiten auf eine noch breitere Basis stellende Items, wie auch mittels non-verbaler Elemente (Bilder, Collagen etc.) sind möglich und sollten verfolgt werden. Dann sind klassische inhome-use-Tests durchaus ein sehr hilfreiches Messinstrument; nicht nur zum abschließenden „Bewerten“, sondern auch als gezielte Hilfe bei weiteren Optimierungen von EdTs und Parfums.



Referenzen

Den an dem *state-of-the-art* der Duftmarktforschung interessierten Leser seien insbesondere die *papers* der ESOMAR-Seminare zu diesem Thema empfohlen: Lyon 1989, London 1991, Köln 1993, Amsterdam 1996.