

Sensorische Analyse: Methoden- überblick und Einsatzbereiche

Teil 7: Difference-from-Control-Test (Teil 2)

<u>Farbbeispiel</u> Nr.			<u>Größe des</u> <u>Unterschieds</u>
1			<u>1</u>
2			<u>3</u>
3			<u>0</u>
4			<u>4</u>
5			<u>2</u>

SENSORISCHE ANALYSE: METHODEN- ÜBERBLICK UND EINSATZBEREICHE – TEIL 7: DIFFERENCE-FROM-CONTROL-TEST (TEIL 2)

Im DLG-Arbeitsblatt 2/2013 wurden bzgl. des Difference-from-Control-Tests Anwendungsbeispiele aus der Qualitätskontrolle vorgestellt. Im folgenden Teil 2 wird der Einsatz des Difference-from-Control-Tests als Alternative zu Unterschiedstests oder Profilprüfungen fokussiert.

3. Einsatz als Alternative zu Unterschiedstests oder Profilprüfungen

3.1 Grundlegender Ablauf

Es werden eine Referenzprobe und eine oder mehrere codierte Proben vorgelegt. Die Probenaufstellung erfolgt im balancierten Design. Die Proben werden den Prüfpersonen zufällig zugewiesen. Die Prüfer beurteilen die Gesamtabweichung oder die Abweichung in vorgegebenen Merkmalen zwischen Referenz- und Prüfproben auf einer Unterschiedsskala. Sie sind dahingehend zu informieren, dass sich unter den Proben auch mit der Referenz identische Proben befinden können.

Für die Auswertung ist zwingend notwendig, dass die Referenzprobe verblindet mit getestet wird. Mittels statistischer Methoden wird bestimmt, ob der Unterschied zwischen Referenz und Prüfprobe(n) signifikant größer als das „Grundrauschen“ des Panels ist. Ergänzend werden Beschreibungen des vorliegenden Unterschiedes erhoben, die der Prüfungsleiter zusammenfasst.

3.2 Prüferqualifikation und -anzahl

Der Difference-from-Control-Test wird meist mit geschulten Prüfern durchgeführt, kann aber auch mit ungeschulten Verkostern, d.h. Konsumentenpanels, angewendet werden. Erstere müssen sich nach DIN 10961 als Prüfer eignen und mit Methode und Skala vertraut sein. Innerhalb eines Panels muss die Prüferqualifikation vergleichbar sein.

MEILGAARD et al. (2007) empfehlen, jede Probe (auch die verblindete Referenzprobe) mindestens 20mal vorzulegen, sofern mit einem geschulten Panel gearbeitet wird. Für ungeschulte Testpersonen wird die Mindestvorlage von 50mal genannt.

Praxiserfahrungen zeigen, dass sich bereits mit acht bis zwölf zuverlässigen Prüfern sinnvolle Ergebnisse erzielen lassen (vorausgesetzt, jeder Prüfer verkostet jede der zu prüfenden Proben). Weniger als acht Prüfer sind nicht zu empfehlen.

3.3 Testbedingungen

Die Schulung der Prüfpersonen wie auch die Durchführung der Difference-from-Control-Tests sollte in einem Prüfraum nach DIN 10962 erfolgen. Die Proben sind zu verblinden und störende Farbeinflüsse gegebenenfalls durch Abdunklung oder Rotlicht auszuschließen. Die Proben werden balanciert und randomisiert gereicht werden. Nach MEILGAARD et al. (2007)

sollte man nach Möglichkeit alle Prüfproben gleichzeitig vorlegen. Zu beachten ist, dass die Erwartungshaltung der Prüfer bei dieser Testmethode die Bewertungen sehr stark beeinflussen kann. Deshalb muss beim Testaufbau so vorgegangen werden, dass mögliche Erwartungshaltungen weitestgehend minimiert werden (siehe Praxisbeispiele 1 und 2).

Die Art der Balancierung hängt von der Zahl der Prüfproben ab. Werden lediglich zwei Proben (Referenz und Alternativprobe) miteinander verglichen, könnte man theoretisch von folgendem Design ausgehen:

Ref (P1) P1 – P2

Ref (P1) P2 – P1

Allerdings läuft man Gefahr, dass die Prüfer, sobald sie entdeckt haben, dass der Testaufbau immer eine identische Probe beinhaltet, einen Duo-Trio-Test absolvieren und das Ergebnis so verfälschen (siehe Praxisbeispiel 1).

Besser ist es, das Design des Same-Different-Tests zu nutzen. Hier sind vier Probenkombinationen möglich:

P1 – P1

P1 – P2

P2 – P1

P2 – P2

Diese müssen in einem Test zu gleichen Anteilen dargeboten werden. Wurde der Difference-from-Control-Test ausgewählt, weil es sich um stark nachhängende Proben handelt, bewertet jeder Prüfer nur ein Probenpaar. Bei unproblematischen Proben kann jeder Prüfer zwei Probenpaare verkosten, die Prüferzahl ist entsprechend zu halbieren. In diesem Fall sind die Prüfer zu informieren, dass ihnen unterschiedliche Kombinationsmöglichkeiten von Probenpaaren vorliegen können:

identisch – identisch

identisch – unterschiedlich

unterschiedlich – identisch

unterschiedlich – unterschiedlich

Sollen zwei oder mehr Proben mit einer Referenzprobe verglichen werden, kann pro Prüfer ein Probenpaar, bestehend aus Referenz und verblindeten Prüfproben, gereicht werden. Bei z. B. drei Proben (1 Referenz, 2 Prüfproben) bestehen folgende Kombinationsmöglichkeiten (vgl. Tab. 3):

Tab. 3: Probendarreichungsmöglichkeiten bei 3 Proben im Set

Kombination	Referenz	verblindete Proben		
1	1	1	2	3
2	1	1	3	2
3	1	2	1	3
4	1	2	3	1
5	1	3	1	2
6	1	3	2	1

Alternativ kann man die Proben auch in Paaren reichen, so dass jeder Prüfer drei Probenpaare erhält. Bei der ersten Probe handelt es sich jeweils um die Referenzprobe, die zweite, verblindete Probe entspricht der Referenz oder einer der mit ihr zu vergleichenden Proben (vgl. Tab. 4):

Tab. 4: Probendarreichungsmöglichkeiten bei 3 Proben, paarweise zur Referenz gereicht

Kombination	Probenpaare		
1	1-1	1-2	1-3
2	1-1	1-3	1-2
3	1-2	1-1	1-3
4	1-2	1-3	1-1
5	1-3	1-1	1-2
6	1-3	1-2	1-1

Die letztere Möglichkeit der Probendarreichung ist v.a. zu empfehlen, wenn es sich um Proben handelt, die kalt oder heiß zu verzehren sind. So wird sichergestellt, dass Referenz- und Vergleichsproben bei gleicher Temperatur verkostet werden.

Die Zahl der Proben, die von einem Prüfer pro Sitzung verkostet werden können, ist abhängig vom Probenmaterial. Sollen mehr Proben beurteilt werden als die Prüfer hintereinander bewältigen können, muss im Blockdesign getestet werden. Bei vier Paarvergleichen oder bei Probensets, bestehend aus Referenz und 5 verblindeten Proben, ist oft das Limit erreicht.

3.4 Unterschiedsskala

Prinzipiell kann man auch für andere Anwendungszwecke die in der Qualitätskontrolle verwendeten Skalen nutzen. Der Vorteil ist, dass Prüfer, die regelmäßig in der Qualitätskontrolle verkosten, die Skala sehr gut kennen und sie gut auf andere Testobjekte übertragen können. Um Unterschiede zwischen den Proben besser herausarbeiten zu können, kann es jedoch sinnvoll sein, längere Skalen zu verwenden. Bei einer Linienskala ist eine gute Abstufung möglich, aber bei einer Kategorienskala, die bis 4 oder 5 geht, lässt sich unter Umständen schwer eine Aussage treffen, welche Variante (besonders wenn jede sehr stark von der Referenz abweicht), die beste ist. Hier sollte man die Skala erweitern, wobei sich sieben Skalenpunkte als praktikabel erwiesen haben. Diese könnte, wie folgt, aussehen:

- (0) kein Unterschied
- (1) sehr kleiner Unterschied
- (2) kleiner bis mittlerer Unterschied
- (3) mittlerer Unterschied
- (4) mittlerer bis großer Unterschied
- (5) großer Unterschied
- (6) sehr großer Unterschied

Skalen mit größerem Umfang sowie die Verwendung von Linienskalen führen zu einer größeren Streuung.

3.5 Training

Der Prüfungsleiter erläutert den Prüfern die Methode und das allgemeine Vorgehen. Er verdeutlicht die Skala anhand von geeigneten Beispielen. Hierfür können Farbbeispiele ge-

nutzt werden, aber auch artgleiche Produkte von verschiedenen Herstellern oder eigens präparierte Proben.

Die während des Trainings erhobenen Daten werden vom Pannelleiter dokumentiert und zur Einschätzung der Prüferleistung genutzt.

3.6 Testdurchführung

Dem Prüfer werden ein oder mehrere Probenpaare oder ein Probenset, bestehend aus Referenz und mehreren codierten Proben, gereicht. Der Prüfungsleiter gibt vor, welche Merkmale (Gesamtunterschied, Unterschied in Aussehen, Geruch, Geschmack...) zu bewerten sind. Je nach Projekt und Ziel des Tests kann er auch genauer werden und Teilbereiche oder bestimmte Merkmalseigenschaften abfragen (Farbe, Oberflächenbeschaffenheit, Konsistenz, Süße...).

Der Prüfer ist über den Testaufbau aufzuklären, um bewusste oder unbewusste Denkprozesse zu minimieren. Er muss bei der Gabe eines Probenpaares wissen, dass es identische oder unterschiedliche Proben beinhalten kann. Bei der Gabe von zwei Probenpaaren muss er über die möglichen Kombinationen (id/id, id/un, un/id, un/un) informiert sein.

Es wird von links nach rechts verkostet, beginnend bei der Referenz.

Es ist ausreichend Neutralisationsmaterial zur Verfügung zu stellen. Bei nicht nachhängendem Probenmaterial kann stilles Wasser oder Mineralwasser bereits ausreichen. Bei glutamat-haltigen Proben kann trockenes Brot oder Matzen das Neutralisieren erleichtern. Bei scharfen Produkten hilft Naturjoghurt, desgleichen bei kühlenden Proben. Bei letzteren ist auch ein schwacher, lauwarmer Schwarztee hilfreich, bei dem im Anschluss mit Wasser gespült werden muss.

Die Daten können per Papierformular (Beispiel siehe Abbildung 3) oder mittels Sensoriksoftware erhoben werden.

3.7 Auswertung

Wurden zur Datensammlung Linienskalen auf Papier verwendet, so müssen die Distanzen mit dem Lineal ausgemessen werden. Wurde eine verbale Kategorienskala verwendet, so müssen den einzelnen Skalenpunkten Zahlen zugewiesen werden, bevor man Mittelwerte und ggf. Signifikanzen berechnen kann. Ergebnisse, die anhand von numerischen Kategorienskalen erhoben worden, können direkt in die Berechnung einfließen.

Abb. 3: Beispiel für ein Formular mit Linienskala

Difference-from-Control-Test - Einzelprotokoll -

Name: _____

Datum: _____

Sie erhalten eine Referenz und mehrere codierte Proben. Bitte verkosten Sie von links nach rechts und geben Sie auf den untenstehenden Skalen an, wie groß der Unterschied zur Referenzprobe ist. Bitte begründen Sie Ihre Entscheidung.

Bitte beachten Sie, dass unter den codierten Proben eine Referenzprobe sein kann. Bitte denken Sie daran, zwischen den Proben zu neutralisieren.

Vielen Dank!

Probe	Geruch	Begründung:
_____	 identisch ----- völlig abweichend 	_____
_____	 identisch ----- völlig abweichend 	_____
_____	 identisch ----- völlig abweichend 	_____

MEILGAARD et al. (2007) geben das Vorgehen zur statistischen Analyse wie folgt an: Wurden lediglich zwei Proben (Standard plus Variante) miteinander verglichen oder sollen mehrere Varianten nur mit der Referenz, nicht aber untereinander verglichen werden, bedient man sich des t-Tests. Sollen multiple Produktvergleiche erfolgen, ist die Varianzanalyse (ANOVA) zu verwenden. Gibt diese Hinweise auf einen signifikanten Unterschied, so bieten sich Dunnett-Test oder Fischer-LSD-Test als post Hoc-Tests an.

O'Mahony (1986) sieht die Analyse mit parametrischen Tests in Zusammenhang mit verbalen oder numerischen Kategoriesskalen kritisch. Skalenendeffekte und die Tatsache, dass die Intervalle zwischen den einzelnen Skalenpunkten unterschiedlich groß sein können, können zu einer Verletzung der Normalverteilung führen. Skalenendeffekte (Vermeidung der Skalenendpunkte durch die Prüfer) können weiterhin dazu führen, dass keine Homoskedastizität (Varianzhomogenität) vorliegt, die ebenfalls Voraussetzung für parametrische Verfahren ist. Sind die Voraussetzungen Normalverteilung und Varianzhomogenität verletzt, erhält man über parametrische Verfahren nur grobe Annäherungen. Daher empfiehlt der Autor einen vorsichtigen Umgang mit den durch parametrische Tests erhaltenen Aussagen zur Signifikanz und rät zu einer Überprüfung mit nichtparametrischen Tests. Für diese müssen weder Normalverteilung noch Varianzhomogenität vorliegen. Vergleicht man lediglich zwei Proben miteinander, verwendet man den Wilcoxon-Test. Bei drei und mehr Proben setzt man den Friedman-Test ein. Beide Tests basieren auf Rangdaten. Die Skalenwerte sind entsprechend zu transformieren. So wird der Probe mit dem geringsten Skalenwert der Rangplatz 1 zugewiesen, der Probe mit dem höchsten Skalenwert der letzte Rangplatz. Weichen die Ergebnisse aus parametrischen und nichtparametrischen Tests voneinander ab, so ist eher letzteren zu vertrauen. Nachteil der nichtparametrischen Verfahren ist, dass man Informationen bezüglich der Größe des Unterschiedes verliert.

Wurden zusätzlich Unterschiedsbeschreibungen erhoben, so sind diese vom Prüfungsleiter zusammenzufassen. Im folgenden Beispiel wurden zwei Rauch-Aromen miteinander verglichen. Der t-Test ergab, dass sich die beiden geprüften Varianten signifikant voneinander unterscheiden. Deshalb wurde näher betrachtet, inwieweit die Prüfprobe von der Referenz abweicht (Tab. 5).

Tab. 5: Beschreibung des Unterschieds zwischen zwei Rauch-Aromen (Auszug)

Prüfer	Unterschiedsbeschreibung
1	rauchiger, stärkere Muskatnote
2	etwas rauchiger
3	stärker, fruchtiger, süßer, mehr nach Kondensmilch, rauchiger
4	deutlich stärkere Nelkennote, mehr nach Schinken, rauchiger
...	
11	stärkere Rauchnote
12	weniger süß, mehr verbrannt, rauchiger

Die Zusammenfassung (Tab. 6) zeigte auf, dass sieben von zwölf Prüfern die Prüfprobe als rauchiger empfanden. Desweiteren wurde sie in drei Fällen als intensiver bzw. nelkiger beschrieben.

Tab. 6: Zusammenfassung der Prüferbemerkungen

Attribut	Schwächer	stärker
rauchig		7
Muskat		1
Intensität		3
fruchtig		1
süß	1	2
Kondensmilch		1
Nelke		3
Schinken		2
frisch		1
Zimt		1
Coumarin		1
würzig		1
stechend		1
verbrannt		1
1x komplett andere Rauchnote (Kondensat)		

Praxisbeispiel 2: Vergleich zweier Weichlakritzproben mit Rohlakritz unterschiedlicher Lieferanten

Ein Süßwarenhersteller musste auf einen anderen Rohlakritzlieferanten umstellen. Um sicherzustellen, dass sich seine Produkte durch den Einsatz einer neuen Rohlakritzqualität nicht sensorisch verändern, bat er um die Durchführung eines Dreieckstest. Da es sich bei Lakritz jedoch um ein stark nachhängendes Produkt handelt, entschied man sich für einen Same-Different-Test. Um dessen Ergebnis wiederum zu überprüfen, wurde zusätzlich die Unterschiedsgröße auf einer Skala von 1 (=kein Unterschied) bis 5 (=sehr großer Unterschied) abgefragt*. Jeder der 39 geschulten Prüfer erhielt zwei verblindete Probenpaare. Die Prüfer wurden informiert, dass es sich bei diesen Paaren um zwei aus identischen oder um zwei aus unterschiedlichen Proben bestehende Paare handeln kann bzw. dass auch ein aus identischen und ein aus unterschiedlichen Proben zusammengesetztes Paar in zufälliger Reihenfolge vorliegen könnten. Im ersten Schritt sollten sie entscheiden, ob die Proben im Probenpaar identisch oder unterschiedlich sind. Im Falle eines wahrgenommenen Unterschiedes sollten die Größe der Abweichung bewertet und der Unterschied beschrieben werden.

Ergebnis und Fazit: Die Prüfer entschieden sich insgesamt 32mal für „identisch“, wobei es sich in 15 Fällen tatsächlich um identische Proben handelte (vgl. Tab. 7). In 17 Fällen waren die Proben im Probenpaar jedoch unterschiedlich. 46mal lautete die Antwort „unterschiedlich“, wobei tatsächlich 24mal identische und 22mal unterschiedliche Proben vorlagen. Der Same-Different-Test wurde mittels χ^2 -Test analysiert, welcher für das Auffinden eines Unterschiedes eine Irrtumswahrscheinlichkeit α von 65 % ermittelte.

Tab. 7: Ergebnisse des Same-Different-Tests und mittlere Unterschiedsbewertung

Proben	Antwort „identisch“	Antwort „unterschiedlich“	Summe	mittlere Unterschiedsgröße
identisch	15	24	39	2,2
unterschiedlich	17	22	39	2,0
Summe	32	46	78	

Ein t-Test bezüglich der Unterschiedsbewertung erübrigt sich, da das identische Probenpaar im Mittel als etwas stärker

abweichend als das unterschiedliche Probenpaar bewertet wurde. Sowohl die Ergebnisse des Same-Different-Tests als auch die Bewertung der Unterschiedsgröße lassen darauf schließen, dass sich die beiden Lakritzproben nicht sicher unterscheiden lassen.

Tab. 8: Getrennte Betrachtung der Ergebnisse der Unterschiedsbewertung mittels Difference-from-Control-Test für die beiden aus identischen (id) oder verschiedenen (un) Proben zusammengesetzten Probenpaare

Probenpaar	Verteilung id/un	Häufigkeit der vergebenen Skalenwerte					mittlere Unterschiedsgröße
		1	2	3	4	5	
1	19x id	3	6	6	3	1	2,6
	20x un	4	7	7	2	-	2,4
2	21x id	10	6	5	-	-	1,8
	18x un	11	4	2	1	-	1,6

Die getrennte Betrachtung beider Probenpaare zeigt verschiedene Effekte auf. Beim ersten Probenpaar entschieden sich im Same-Different-Test überproportional viele Prüfer (ca. 80 %) für die Option „unterschiedlich“, unabhängig davon ob die Proben identisch oder gleich waren. Als Ursachen kommen hier zwei Möglichkeiten in Betracht:

Die menschliche Psyche: Motivierte Prüfer auf der Suche nach Unterschieden „finden“ diese in der Regel, egal ob tatsächlich vorhanden oder nicht. Die Erwartung eines Unterschiedes führt in der Regel zur Bestätigung desselben.

Physiologische Aspekte: Bei Proben, die stark nachhängen, reichern sich bestimmte Bestandteile des Lebensmittels (hier z. B. das Anethol) in der Mundschleimhaut an. Sie sind schwer zu neutralisieren. Trotz gründlichen Mundausspülens kann es so bei der Prüfung einer weiteren Probe zu einer veränderten, d.h. intensiveren Wahrnehmung kommen.

Beim Probenpaar 2 ist die Verteilung der identisch/unterschiedlich-Antworten für die aus identischen Proben zusammengesetzten Probenpaare nahezu ausgewogen. Die aus unterschiedlichen Proben bestehenden Probenpaare wurden von ca. zwei Dritteln der Prüfer als gleich empfunden. Auch hier sind wieder psychologische wie physiologische Aspekte als Ursachen anzusehen:

Die menschliche Psyche: Wurde das erste Probenpaar als unterschiedlich empfunden, so rechnet der Prüfer hier eher mit einem identischen Probenpaar. Um diesem Denken vorzubeugen, wurden die Prüfer darauf hingewiesen, dass auch 2x ein identisches oder 2x ein unterschiedliches Paar vorliegen kann. Ganz ausschließen lässt sich dieser Einflussfaktor aber nicht.

Physiologische Aspekte: Die Mundhöhle der Prüfer war bei der zweiten Verkostung derart belegt, dass sie Probleme hatten, Unterschiede wahrzunehmen. Aus dem Feedback der Prüfer nach erfolgter Verkostung lässt sich entnehmen, dass dieser Faktor den größeren Einfluss besitzt.

Insbesondere letztere Erfahrung zeigt, dass bei den vorliegenden Proben ein Test, der mehrfache Probenvergleiche erfordert, nicht geeignet ist und auch grundsätzlich nur ein Probenpaar

gereicht werden sollte. Um keine falschen Schlüsse zu ziehen, wurden die Ergebnisse des Same-Different-Tests für das erste Probenpaar separat statistisch analysiert. Die Sensoriksoftware Fizz berechnete mittels χ^2 -Test (unter Berücksichtigung der Yates-Korrektur) eine Irrtumswahrscheinlichkeit von 75 %, was wiederum ein nicht signifikantes Ergebnis bedeutet. Dies korrespondiert mit dem Ergebnis des Difference-from-Control-Tests, bei dem die Größe des Unterschieds für das aus identischen Proben zusammengesetzten Probenpaars im Mittel mit 2,6 und für das aus unterschiedlichen Proben bestehenden Paares mit 2,4 angegeben wurde.

4. Fazit

Der Difference-from-Control-Test ist ein vielseitig einsetzbares Verfahren, das Auskunft über Größe und Art des Unterschiedes zwischen zwei oder mehr Proben gibt. Vorteil der Methode ist, dass sich in sehr kurzer Zeit mit einer überschaubaren Prüferzahl aussagekräftige Ergebnisse erheben lassen. Desweiteren lässt sich der Test sehr gut bei nachhängenden Proben oder einem sehr hohen Probenumfang einsetzen. Nachteilig ist, dass die Methode sehr fehleranfällig ist. Im Testaufbau ist unbedingt zu beachten, dass Erwartungshaltungen seitens des Prüfers sehr großen Einfluss nehmen können und durch Balancierung, Information etc. weitestgehend ausgeschaltet werden müssen.

Eine reelle Einschätzung des Unterschieds erhält man erst, wenn man die Referenz verblindet mitlaufen lässt, da zwei Produkte, vergleichend verkostet, in aller Regel unterschiedlich erscheinen. Ursachen sind physiologische wie psychische Gegebenheiten, die zum sogenannten „Grundrauschen“ des Panels führen. Die Frage ist, ob der Unterschied zwischen Referenz und Prüfgut tatsächlich dieses Grundrauschen signifikant übersteigt.

Zur statistischen Analyse können bei zwei Proben der t-Test oder beim Vergleich mehrerer Proben eine Varianzanalyse durchgeführt werden. Deren Ergebnisse sollten dann aber mit nichtparametrischen Tests (Wilcoxon-Test bzw. Friedman-Test) abgesichert werden.

Wird der Test für Fragen der Qualitätskontrolle genutzt und es besteht Unsicherheit, ob Proben tatsächlich freigegeben werden können, muss dies durch weitere Tests abgesichert werden. Dies kann unter Einsatz von mehr Prüfern und einem balancierten Design durch das gleiche Testverfahren oder durch einen klassischen Unterschiedstest (z. B. Paarvergleich, Duo-Trio-Test) geschehen.

Autorin: Ines Strobl, Dipl.-Oec.troph. (FH), Senior Sensory Manager; Silesia G. Hanke GmbH & Co.KG, Neuss, E-Mail: I.Strobl@silesia.com

Weiterführende Literatur:
siehe DLG Expertenwissen 2/2013.



DLG e.V., Ausschuss Sensorik

Eschborner Landstraße 122, 60489 Frankfurt am Main

Telefon: 069/24788-360, Fax: 069/24788-8360

E-Mail: B.Schneider@DLG.org; Internet: www.DLG.org/sensorikausschuss.html