cmj ctj

Unilever

Final Stage Product Testing Guidelines

Issued By PECLT

STRICTLY CONFIDENTIAL
To be used for Unilever Studies ONLY

Updated Q1 2016

The guidelines prescribed in this document MUST be followed for all Final Product Tests.

Exceptions, if any, need to be authorised by the CMI Category VP or CTI Director.

Any category specific protocol will need to comply with the Unilever guidelines.

Table of Contents

1. When to Conduct Product Testing

There are 4 key stages when product testing is done:

Stage 1	**During the 'Ideas' phase of the Innovation Funnel (or the early part of 'Feasibility'), when the research is exploratory in nature**
Stage 2	**During 'Feasibility' when you are trying to select a winning formulation(s) to go into a Final Acceptability Test i.e. expert testing**
Stage 3	**Towards the end of 'Feasibility', when you conduct a Final Acceptability test to ensure the product has sufficient consumer appeal**
Stage 4	**At a point in time unrelated to the Innovation Funnel, when you want to assess consumer reactions to your product versus competitor products**

These guidelines cover Stages 3 and 4 i.e. final product testing pre or post launch (either including or not including competitor products) for *both CTI and CMI*.

A Final Acceptability Test MUST be conducted in the following circumstances:

- **When making a significant change to an existing product (e.g., change to formulation/sensory property/processing etc)**

- **When introducing a new product/brand**

When conducting early stage product development testing (Stage 1 and 2), CTI professionals need to be aware of these guidelines to ensure testing continuity and consistency.

Although all elements of the mix are important, the long term success of brands in the marketplace is absolutely dependent upon **product performance**. Consumers expect products to do a good job, time after time and to live up to the promises our brands make. This kind of continuing satisfaction leads to healthy repeat purchase patterns and this sustains brands in the marketplace over the long term.

Conversely, the most creative Marketing approach will not sustain a brand where product performance is sub-standard. Also, new brands with product performance weaknesses may generate good trial but are likely to fail because of poor repeat rates. The Benchmarking initiative is focused on determining exactly how in market Unilever products perform compared to competition.

Particular care should be taken with Product Rationalisation changes to existing products designed to save money on existing brands. These clearly run the risk of downgrading product performance and alienating your existing buyer base. As this can take a considerable period of time to manifest itself in lost sales, it is often too late for some brands to recover by the time such performance problems are detected and addressed.

For these reasons, no significant change should be made to an existing brand and no new brand should be launched before it has passed **a full-scale product test,** drawn from a representative sample of relevant consumers and used under normal usage conditions.

Note 1: The basic product test methodology outlined in these guidelines can also be used to evaluate structural paradigms which affect the in use characteristics.

Note 2: Product testing is NOT the appropriate method to use when the objective is to estimate the relationship between product performance and pricing or sales. In these cases a Simulated Test Market approach should be used. STM approaches are covered in separate guidelines.

2. The 4 Basic Types of 'Problem' for which Final Product Testing is Used

The table overleaf summarises the key features of the different types of product test covered by these guidelines.

Problem Type	Issue Addressed	Test Design	Branded vs. Unbranded	Control / Benchmark	Action standards	Sample Structure
1. New launch – (new brand or line extension)	Consumer acceptability of new product or brand	Monadic concept / product test	Branded with concept/ Branded only	Estimated source of volume ie. A leading competitor OR current variant in the range	Significantly superior on at least one OO or key attribute measure at 95% LoC t-test 2-tailed. 80% 1-tail test when parity check is requested	Category users (N=200)*
2. Existing product improvement	Consumer acceptability of product improvement	Monadic product test	Branded (with new concept if appropriate)	Current formulation (with new concept if a concept is used for test product)	Significantly superior on at least one AS measure at 95% LoC t-test 2-tailed. 80% 1-tail test when parity check is requested	Brand users + non-rejectors (N=200) with boost to achieve 100 heavy brand users Optional boost up to 100 of each user group. If parity performance may be acceptable for launch, then a parity check among 150-200 brand users is required
3.VIP - Product rationalisation/ Harmonisation/ Inter changeability/ Flexibility	Consumer acceptability of product change	Comparative test	Branded	Current (or previous current to avoid salami slicing)	Parity to current at 80% LoC (see section 10- Action Standards)	Brand Users (split Heavy/Medium/Light users, based on volume split) Interchangeability/ Flexibility (n= 400) Rationalisation (n= 250) Harmonisation (n= 200)
4. Competitor Benchmarking	Consumer acceptability of our products compared with the competition	Monadic OR Sequential Monadic	Unbranded	**Competitor**	Significantly superior on Overall Opinion or all key attributes at 95% LoC t-test 2-tailed on combined position scores (unless an order effect is found)	Category users (N=200)* with boosts for users of each brand tested (100+) if required. Categories might choose to split the sample equally among users and non-users

* When there are no subgroups to be analysed this can be reduced to a minimum of 150

3. Test Design: Monadic vs. Comparative

The decision to be made here is between sensitivity and realism.

The more realistic monadic test design is used for most situations – for New Products, Existing Product Improvements and Competitor Benchmarking.

The exception is VIP which requires the more sensitive comparative test.

A good product test needs to strike a balance between realism and sensitivity of measurement. A test design that calls for consumers to use a product under highly artificial conditions is likely to yield distorted results. On the other hand, in certain cases, some degree of lack of realism may be necessary if respondents are to notice and respond meaningfully to the difference being tested.

The main choice to be made is between a **monadic design** where each respondent tests only one product and a **comparative design** where the respondent tests two or more products.

The **monadic design is more realistic** because in real life consumers normally use one product at a time and compare them (in their minds) to other products they have used.

If product differences occur in a monadic test, then it is likely that they would also be picked up in the marketplace. If differences do NOT occur monadically, then it is unlikely that they would be picked up in the marketplace and generate any differences in sales.

This is a critical factor in all types of test where monadic testing is recommended:

New Product/ Brand Launch – you want to be sure that any new product launched will be of good enough quality to succeed before you go to the expense of launching it.

Existing Product Improvement – you want to be sure that any money you invest in improving the product will be well spent (in terms of generating greater consumer interest/satisfaction).

Competitor Benchmarking – you want to be sure that any differences you observe between our products and our competitors are meaningful and not exaggerated.

In a typical monadic test, the total sample is divided into matched cells, each cell being given a single product to be used for a fixed period of time. After this usage period, consumers in each cell are interviewed to ascertain their opinions of the product. In some cases, assessments can be captured via the use of a self-completion questionnaire.

The **comparative design is more sensitive** and is more likely to pick up differences, because the respondent is being asked to directly compare and contrast the two products they are testing.

If product differences occur in a comparative test, it does not necessarily mean that they would also be picked up in the marketplace, but it does highlight some degree of risk. However, if they do NOT occur comparatively, you can be pretty sure that they won't be picked up in the marketplace and generate any loss in sales.

This is a critical factor in VIP tests where you want to be sure that any changes you have made to an existing product will NOT be noticed by consumers.

However, there are some situations where a comparative test is not always practical. In these cases, a monadic test design should be used, with a larger sample size to increase sensitivity:

- where the effects of the first product can be expected to carry over to the second product e.g., anti-dandruff shampoo, whitening toothpaste.

3.1 Types of Comparative Tests

There are two types of comparative tests
1. Paired comparison – For VIP
2. Sequential monadic – For Star Wars benchmarking

Paired Comparison (for VIP): provides overall opinion and preference ratings. This is mandated for Final decision making for all VIP initiatives and should be used as a Go/No Go metric.

Sequential Monadic: provides preference ratings, diagnostics from product shown in each (first, second etc.) position. If it is important that detailed attribute data is provided for products in all positions, as well as overall preference, the more comprehensive sequential monadic design should be used for diagnostic understanding.

In a comparative test the key measure is the preference question.

In a **paired comparison test for VIP** the respondent tests the first product and a recall interview is conducted to only provide an Overall Opinion rating of this product. They then test a second product and at the second recall interview provide an overall opinion rating of the second product as well as being asked which of the two products was preferred, plus preference on key attributes.

In a **sequential monadic** test the respondent is given the first product to test and a recall interview is conducted to provide ratings of this product on a range of characteristics. They are then given the second product to test and at the second recall interview they provide ratings of this second product as well as being asked which of the two products is preferred.

In both comparative designs it is important that an equal number of respondents test each product first and second to balance order of testing effects which frequently occur in comparative tests.

3.1.1 Analysis of Sequential Monadic Tests

Data from a sequential design is best analysed taking explicit account of order of presentation or context. The evaluations of a product seen in the second position are as insightful as those from the first position: the difference in ratings is influenced by the presence or absence of other products.

3.1.2 Tests Involving More than One Test Product

If there are more than two test products being tested in a comparative design as alternatives to current, the test design should be:

Cell 1: Test A vs. Current
Cell 2: Test B vs. Current
Etc.

A round-robin design, where the test products are also tested against another, is NOT recommended for a final product rationalisation test, as the key question to answer is how the test products compare with current, not with one another.

4. Branded vs. Unbranded

Branded product tests provide a meaningful context for respondents to evaluate the product. Blind tests are appropriate for:

- **early stage product testing where the emphasis is on isolating the impact of new technology (early product testing not covered in this document)**
- **competitor benchmarking when the focus is on the product (and not on the concept or claim)**

Consumers ultimately buy brands not formulations. There are well-documented cases in which responses to branded products are quite different from responses to products where brand identification has been removed.

In a blind competitor benchmarking test, all products should ideally be repacked in plain packaging. In such cases, a question should be included in the interview asking respondents whether they recognized the brand, and if so, which brand they thought they were evaluating testing.

Final test should always be branded (including fragrance and flavor tests). If the test is a branded test, but the packaging for the new product is not final, then all products (including the control) should be placed in unbranded packaging, but with a photograph of the final finished pack. Every effort should be made to test the product in its finished form including the final packaging as consumers would evaluate the holistic product design including the packaging.

In the case where a competitive product is used as a control in a test of our own brand, the principle we should use is that we should NOT compromise the basic principle of testing branded, simply to accommodate a competitive control. This leaves us with 2 options:

- run two tests (the first branded within our products, the second unbranded of the winner against the competitor);

- run a concept product test with each product branded as itself

When a test that would normally be branded has been conducted unbranded (to maximise sensitivity at an early stage of development) a branded test must be conducted subsequently.

Regardless of whether testing branded or unbranded, with or without concept, **detailed usage / cooking instructions should be provided where appropriate, together with any required advice/ warnings related to product usage.**

In circumstances where it is difficult to “de brand” products (e.g the shape of a Cornetto is unmistakable) , and the test is still conducted unbranded you must include a brand recognition question and use the wording and scale at the end of the questionnaire in Appendix 1. For analysis purposes look at subgroups of those who recognised the product and those who did not.

5. Use of Control Product

A control product should be included in the test and used for setting action standards.

For Existing Product Improvement and Existing Product Rationalisation, control will normally be the current formulation. For PQB Benchmarking, the control will be the competitor product.

- **The best selling variant ("biggest share") should be used if the whole line cannot be tested due to cost constraints.**
- **Control products will usually differ by country dependent upon category**

For New launches it will depend on the nature of the launch.

- **Line extension replacing an existing variant – Variant being replaced should be used as the control**
- **Line extension adding a variant – Largest selling variant in existing range should be used as the control**
- **New brand launches – Competitor which would be the source of growth should be used as the control**

In cases where it is not possible to identify a control then RED can be used.

For cases where an existing product is being improved, the current product should be used as the control.

If the new formulation is to be tested under a new concept, then the same should be done for the current product. This is so that we can isolate the product effect.
Note: In a subsequent mix test, the new and current formulations would be tested under their respective concepts to determine whether the new mix beats the current mix.

For New Product Final Acceptability Tests the decision depends on whether the product is for a new brand or a line extension of an existing brand.

For a line extension there could be two scenarios

1. We are adding an additional variant – here the control should be the largest selling existing variant in the range, or a competitor if more suitable
2. We are replacing an existing variant with a new one- here the control should be the variant being proposed to be replaced

For a new brand, **the control should be the competitor**, which is likely to be source of volume.

CAUTIONARY NOTE: In a situation where several monadic tests have been conducted earlier, it can be tempting to try and compare the results of your latest test with these scores directly, thus avoiding the need for a control leg. However, experience has shown that other factors can affect the results you get (e.g., weather, interviewer effect, competitor activity, market dynamics etc) and adversely affect comparability. Hence, it is not recommended to use data from past tests but to include an additional leg in the new study.

For Product Rationalisation, the control will be the current formulation.
It is worth thinking about whether you are in danger of ‘salami slicing’ your product. This term refers to situations where product quality was consistently downgraded over time. Each individual downgrade was not picked up by consumers, but if you were to test the product you have now against the one you had a few years ago, you would find big differences.

If you think this could be the case, then you should discuss with Marketing and R&D whether to include a benchmark product in the test that represents the quality of the product at an earlier point in time.

For Competitor Benchmarking (PQB), the control/s will be a competitor
The choice of the competitor/s should be defined in the Brand Performance Standards (BPS) of the brand. We need to cover all category users. However, each category needs to decide if it wishes to split the sample equally among users and non-users.

6. Sample Structure and Composition

6.1. Who to Include In the Test

The study objectives will determine the appropriate sample, but in all cases, the key question to ask is "which group or groups of people is this product seeking to address or influence?".

For a New Product/Brand the sample should be a general sample of category users or if a new category entry, non-rejectors of the concept (defined as top 3 box acceptors on the purchase intention scale)**. This is because you cannot predetermine who will or won't be interested in the new brand.**

For Existing Product Improvements, the test should be conducted amongst product/brand users and non-rejectors. Heavy brand users should be augmented to reach a total of 100 heavy brand users per product.

For Existing Product Rationalisations, the test should be conducted only amongst brand users in order to make sure you do not jeopardise your current sales. The sample (consisting of brand users) should be split into heavy, medium and light users in proportion to the contribution to the volume of the SKU / variant which will be impacted. The data on volume contribution by heavy, medium and light should be obtained from the Consumer Panel.

For Competitor Benchmarking, the test should be conducted amongst a general sample of category users, with boosts of relevant brand users. This is so that you can get a broad category context as well as understanding your own / your competitors' consumers.

For most product testing involving categories in which Unilever is involved, **women** are more likely than men to be test respondents as they are the primary users as well as the "deciders" of what products to buy.

However, this is not always the case since, as for several categories and brands, men are active users and either directly make purchases or significantly influence the purchase decision.

Whether the test sample is composed exclusively of women or includes men (or for that matter children) is a decision that must be made based on available information about the importance of each subgroup to product use and purchase decision/influence.

6.2. Representativeness

Once the sample has been defined, every effort must be made to recruit a **representative sample** of those people. For existing brands, the sample should cover at least 70% of the consumption of the brand.

Too often, samples are biased towards those sub-groups of category users who are either most accessible or co-operative. Without safeguards, it is easy, for example, to over-sample women over 50 who are more likely to be found near test centres and who are more likely to have the time to participate in a study.
Note: Setting quotas on housewives usually helps to avoid this phenomenon.

However, equally unacceptable is a sample that has been too restrictively defined, for example, women aged 18-49 if in fact women over 50 represent a significant segment of category or brand usage.

For representative sampling to be achieved, it is necessary to set specific quotas and to balance the composition of the sample across the cells at the analysis stage.

Care must also be taken to ensure geographical representativeness within a country. This should also consider city tiers and urban-rural divides.

6.3 Security (Who to exclude from the Test)

Because of the risk of early prototypes becoming known to competitors, fieldwork should avoid areas where competitors have offices/factories. Each country should prepare, by category, a list of areas to be avoided, and make it available to their market research agencies conducting product testing work.

Similarly, consumers who could be security risks should be excluded from the study sample. In particular, people who work, or who have family or relatives who work in the following areas should be excluded:

- Advertising
- Product Manufacture of (insert category) products
- New Product Development
- Design Agency
- Perfume House
- Construction
- Military
- None of these

In cases where security risk is particularly high, excess product or empty packaging should be collected from consumers after testing.

See Appendix 3 ‘Confidentiality Agreement’.

7. Sample Size

The minimum sample size is 200 respondents per cell. When there are no subgroups to be analysed this can be reduced to a minimum of 150.
The minimum sample size for subgroups to be analysed separately is 100.

Larger sample sizes are mandated for specific Product Rationalisation Tests. Larger sample sizes should also be considered for Product Improvement where action standards among sub-groups need to be met.

While a large sample size increases the likelihood of finding a significant difference, it clearly costs more. However, a test where sample size is too small to detect a meaningful difference would be a complete waste of money.

A sample size of 200 represents the optimum balance between cost efficiency and sensitivity.

For Existing Product Improvements, heavy brand users should be augmented to reach a total of 100 heavy brand users per product.

Note: Under conditions where relatively large differences are anticipated with some confidence, sample size can be reduced to 150 respondents. However, this should be seen as the absolute minimum for final product tests and in such cases no sub-group analysis should be conducted.

8. Cell-to-Cell Balancing

When more than one product is being tested, each of the consumer cells in the test must be matched in terms of key variables that may influence product appreciation.
These variables do impact incidence levels and would include demographics (age, LSM, presence of children etc.) and other factors which would influence product evaluation e.g. water hardness for Laundry, hair type for shampoo. If there is a significant imbalance in characteristics across test cells, this can invalidate the comparability of test results.

Quota sampling approaches should be used to achieve reasonable test cell balance. In cases where proper balance is not achieved in sampling, then post-test re-weighting of results may be required to assure that reliable cell-to-cell comparisons can be made.

NOTE: It is vital that each interviewer places a similar proportion of ALL products in the test, as interviewer bias can be significant.

9. Use of Concept

For New Products/Brands, a concept should be shown to the respondents.

For improvements to an existing brand, the decision on whether or not to include a concept will depend both on whether it is intended to communicate the improvement (or not) as well as the specific test objectives.

Including a concept will enable you to see whether the product lives up to the expectations generated by the concept. However, you may not wish to 'prompt' respondents with a concept but simply observe what, if any, differences they notice.

The same concept should be used for all products in the test of an existing brand so that the product effect can be isolated (same concept for the test product and control/current product). Only where the change in formulation requires a change in concept, can the concepts differ.

Respondents should be allowed to view the concept throughout the concept evaluation interview. In general, respondents should not be allowed to retain the concept during the phase of product usage and product recall interview because the focus should be on the product.

For Existing Product Rationalisation, a concept should NOT be used since there will be no communication of the product change.

In cases where a competitor concept is to be included, it is the responsibility of the brand team to create this, bearing in mind that it must be at a comparable level of quality to that of the Unilever concept to avoid creating any bias. A few checks should be applied to ensure that the competitor concept is not being unfairly treated:

- Number of words on the concept should be similar to Unilever concept – variation of up to 20% is acceptable
- Similar number of products should be featured on the concept
- Quality of the photographs should be comparable

10. Action Standards

Action Standards must be agreed by all parties prior to the test and must be written down as part of the brief and Market Research proposal.

Action Standards should include a measure of overall opinion as well as 2-3 key attributes.

Action Standards must specify the statistical test of significance to be used and the level of significant difference required versus the control product. The statistical test used should be applied to the mean score (as opposed to 'Top box' type scores).

Thought should be given as to whether action standards should be set on total sample and/or within specific sub-samples (e.g., loyal users versus occasional users, users versus non-users). For Product Improvements and VIPs a parity check at 80% 1-tail should be conducted among heavy brand users for diagnostic purposes.

If there is more than one measure in the action standard, refer to whether the action standard should be AND or OR.

The action standard MUST always be set vs. a benchmark and NOT vs expectations set by concept.

Action standards should be set on those questions that are really critical, rather than on a wide variety of measures. This should include:

- Overall Opinion

- Key Attributes (2-3 in number) – The attributes must be a combination of
  - Attributes included in the BPS
  - Attributes that the product change seeks to influence

The number of key attributes should be no more than 2-3.

Action Standards must specify the statistical test of significance to be used and the level of significant difference required versus the control product. The statistical test used should be applied to the mean score (as opposed to 'Top box' type scores). Means should be based on all answering. When Action Standards are reported, the base sizes and statistical test and result should be indicated on the chart.

Action standards must not be set on Purchase Interest which is not to be included in the questionnaire. Action standards must not be set on product ratings matching concept ratings (product meeting concept expectations).

Action standards will vary according to the type of 'problem' the test has been designed to answer and, therefore, it is impossible to state in this document exactly what they should be. However, here are some of the key issues to consider:

With an Existing Product Improvement:
We typically would want the test product to be rated (significantly) higher than current. This is because there is usually a substantial cost involved in changing the product e.g., R&D costs in developing the product, additional formulation costs, supply chain costs in changing over manufacture from one to the other etc. So in these types of test we want to make sure that the change is 'worth it'.

In the event that the business decides to go ahead despite parity performance to current a parity check at 80% among heavy brand users must be done to ensure that there are no indications of inferiority relative to the current product amongst current user base.

With an Existing Product Rationalisation (VIP):
We would normally want the test product to be rated the same as / not worse than current (also known as 'parity'). This is because we do not want to alienate our existing buyers by introducing a product that is worse than the one they are currently using.

In VIP testing there are 3 levels of action standards – preference, overall opinion and alienation. It is important that all 3 of these are met. For each of these the significance level is determined by the kind of VIP i.e whether it is a flexibility, interchangeability or rationalization. (See Analysis of Results and Significance Testing section for more details on significance testing – Section 11). For VIPs a parity check at 80% 1-tail should be conducted among heavy brand users for diagnostic purposes.

With Competitor Benchmarking, the action standards should be guided by the Brand Performance Standards. The BPS should distinguish between attributes that are category "must have's" and attributes that are "differentiators" for the brand. The action standards should be:

a) Superiority on differentiators
b) At least parity on must haves

With New Products: The action standards should be guided by the business scenario

- Replacement of an existing variant - you would expect the new variant to be rated significantly higher than the variant it is replacing.
- Introduction of a new variant to an existing range – you would expect the new variant to be rated significantly higher than the best-selling variant in the range
- Introduction of a new brand – you would expect the new brand to be rated significantly higher than the competitor which would be source of volume.

Action standards set out the actions which we will take under different outcomes of the research and they ensure that we think through carefully what we intend to do with the results before we commission the test.

If there is more than one measure in the action standard, refer to whether the action standard should be AND or OR. Having too many attributes with an 'and' condition is making the action standards very strict. Having too many attributes with a 'or' condition is making the action standards not very strict. In addition, in the case of action standards that include multiple measures, a decision tree approach should be considered.

And finally, action standards should not be vague and non-committal (e.g., "results will be used to decide what action to take"), unless the research is purely exploratory.

To aid in reporting, templates can be found in the appendix for:

- Action Standards (Appendix 4)
- A one-page report summary called the “Power Page” (Appendix 5)

11. Analysis of Results and Significance Testing

Types of significance tests

Two-tailed test:
When there is no prior hypothesis which is better i.e., when the action standard does not focus on whether one alternative wins (or loses).

One-tailed test:
When there is a prior hypothesis that one is better/not worse
i.e., when the action standard focuses on whether one alternative wins (or loses)

For most product tests – New Product/Brand, Existing Product Improvement and Competitor Benchmarking - the 95% significance level should be used (based on a 2 tailed test). A check for inferiority at 80% 1-tail can be used to check among heavy brand users.

For Existing Product Rationalisation, where you are trying to test for parity, and where 80% significance level should be used (based on a 1 tailed test (Rationalisation/Harmonisation), a 2 tailed test (Flexibility/Inter changeability) using preference data.

In all cases except alienation and preference, the statistical tests should be applied to mean score data rather than Top/Top 2 boxes.

Statistical Tests for Significance:

As stated above, for Existing Product Improvement Tests, New Product Tests and Competitor Benchmarking, the 95% confidence level should be used (based on a 2 tailed test) for key action standards. The reason for this is that it gives you greater confidence that consumers will notice the difference in the marketplace.

The 90% confidence level can be used to identify 'directional' differences. Directional differences should only be used for additional guidance and not part of an action standard.

Typically the action standards for a product change should be based on superiority at a 95% confidence level among the total sample. One tailed 80% testing should be conducted among heavy brand users in cases where the business decides to go ahead despite the lack of superiority. To ensure that there are no indications of inferiority relative to the current product. If no superiority is set and overall action standards are set to parity then the **80% 1-tail test** should be used. It is strongly recommended that parity action standard be NOT set for cases that involve a change to an existing product.

To conduct an 80% 1-tailed test within a product improvement (renovation) project each of the following criteria must be met:

a) Unilever to determine if a study falls under renovation at briefing stage.

b) Size of the main (general) sample must be between 200 and 250. Deviating from these recommendations will impact the sensitivity of the statistical tests.
    I. There should be at least 100 heavy brand users per cell.

c) A pre agreed list (by Unilever) of 1-3 attributes + overall Liking should be considered for the test and 80% testing will only be applied to these. Significant increases in the number of instances that the 80% 1-tailed test is applied will increase the likelihood of a false read and hence a potential incorrect conclusion. By exception the CMI VP could authorise up to 6 attributes to be subjected to the 80% 1-tailed test but no more. IPSOS will not provide 80% 1-tailed tests on subgroups or attributes that are not pre-agreed.

The higher the confidence level the less likely a difference flagged as significant will happen by chance (i.e. a false positive error).

For Product Rationalisation Tests and Benchmarking tests which use a sequential monadic design the three point preference scale shown below should be used.

I preferred the product I tried first
I preferred the product I tried second
I have no preference

For analysis purposes the responses to the no preference should split evenly between the test and current product.

Statistical testing for Benchmark tests

For Competitor Benchmarking, Unilever mandates the 95% confidence level (two-tailed test) t-test for key action standards. The reason for this is that it establishes clear likelihood that consumers will notice the difference in the marketplace.
The combined 1st and 2nd position data (aggregated sequential results) will be reported for studies where there are no order effects.

Order effects are defined as situations where people may judge the second product relative to the first one rather than an internal norm.
To ensure that any potential order effect - if at all present - is not big enough to cause a change in conclusion, the following steps are required:

1. Compute the difference between the Unilever product and a competitive product in the First position. Call that difference (A).
2. Then compute the difference between the Unilever product and a competitive product in Total. Call that difference (B).
3. If A and B have different signs (one is positive and one is negative) then one should use the test results on First position only (reporting monadic data). Otherwise, we conclude report Total results (aggregate sequential-monadic data).

Note that this is an effect that needs to be measured on both Overall Opinion AND Key Attributes and needs to be measured for each test. Indicate on the Action Standard chart if an order effect has been found and whether combined or first position results are in the Action Standard using one of the following notes:

1) Note: No order effect was found so the combined results are reported
2) Note: An order effect was found so the monadic results are reported

Using the aggregated data set rather than only the monadic read would strengthen the reliability of the results (as the number of consumers testing a product is doubled, reducing the likelihood of a Type 2 error).

Statistical testing for VIP tests

With an Existing Product Rationalisation, we would want the test product to be rated the same as/not worse than current (also known as ’parity’). This is because we do not want to alienate our existing buyers by introducing a product that is worse than the one they are currently using.

In VIP testing there are 3 levels of action standards – preference, overall opinion and alienation. It is important that all 3 of these are met. For each of these the significance level is determined by the kind of VIP i.e whether it is a flexibility, interchangeability or rationalisation.

i. Preference for VIP testing

For **Flexibility and Interchangeability**, we are looking for *no significant **difference** between the % of consumers preferring current and those preferring the VIP*.
Statistical test – 2-tail, z-test at 80% level of confidence

For **Rationalisation and Harmonisation** we are looking for *no significant **loss** on the % of consumers preferring the VIP compared to those preferring the current. Statistical test – 1-tail, z-test at 80% level of confidence*
(this test is only to ensure not significantly worse, it MUST not be used to look at significant improvement)

ii. **Monadic – Overall Opinion for VIP testing**
(Action Standard for Overall Opinion needs to use combined 1st and 2nd position data. In addition, a check should be done on the 1st position monadic data to ensure the conclusion is consistent with the combined position data. If results differ report Action Standards on combined data and include a note on Action Standard chart with the results from 1st position. If 1st and combined position results do not conflict note in the Action Standard Chart that 1st position results were checked.

For **Flexibility and Interchangeability**, we are looking for *no significant **difference** between the mean score of current and the VIP.*
Statistical test – 2-tail, t-test at 80% level of confidence

For Rationalisation and Harmonisation we are looking for *no significant loss on mean score on the VIP compared to the current Statistical test – 1-tail, t-test at 80% level of confidence***iii. Alienation for VIP testing**

Bottom 3 or 4 Box (B3B/B4B) counts on Overall Opinion (as percent of total sample) for those preferring Current and rating VIP in B3B/B4B should not be significantly larger than B3B/B4B counts of Current for those preferring VIP - 1 tail , z- test @ 90% level of confidence. See below for when to use Bottom 3 versus Bottom 4 Boxes

• Bottom **4** box Overall Opinion for the following regions or countries:

□Africa
□China
□India
□LATAM
□Middle East

• Bottom **3** box Overall Opinion for the following regions:

□Asia (except China and India)
□Europe
□North America

If alienation of the VIP product is less than or equal to 2% (not rounded), then no statistical testing is needed. The alienation test is passed. If alienation is greater than 2% (not rounded), then a 90% 1-tailed z-test is calculated. If the test is not significant then the alienation test is passed. If the test is significant then the alienation test is not passed. Standard statistical tests will be applied to mean score and Top/Top 2 Box (as appropriate) data on presentation charts and topline documents using

Significant at 80%
Significant at 95%

Summary of statistical analysis guidelines:

Test Types	**Description**	**Final Stage Analysis**
New Product	Comparison of variant to control where analysis should support superiority or whether or not a variant is worse than control	T-Test (Two-tail) for superiority and 80% 1-tail test when parity check is requested
Product Improvement	Comparison of variant to control where analysis should support superiority or whether or not a variant is worse than control	T-Test (Two-tail) for superiority and 80% 1-tail test among heavy brand users for parity check
Benchmarking	Comparison of variant to control where analysis should support superiority or whether or not a variant is worse than control	T-Test (two-tail) for superiority
Product Rationalisation	Comparison of variant to control where analysis should support not significantly worse	T –test, Z-test (One/two tail 80% or 90% confidence level as detailed above)

As well as reporting the key measures in total, an investigation of the relationship between the measures should be undertaken. This can be done using a cross-tabulation of the key product measures (i.e. important individual attribute ratings) by such critical measures as overall opinion and brand usage.

The distribution of the responses on key scales as well as the mean scores should be examined, though not necessarily presented.
A template for summarising results against action standards can be found in Appendix 4 - *Action Standards Template*.

12. Where to Conduct Product Tests

Product tests should involve consumers using the test product in a 'normal' consumption situation. For most Unilever products this will be in the home.

A central location test should be considered only if it replicates normal consumption and it should be agreed with the VP CMI of the category.

The objective of quantitative product tests is to determine how consumers are likely to respond to product change in the "real world", under normal use conditions. For most Unilever products this means that tests should be conducted **in-home**.

For some Food products where consumption occurs outside the home, Central Location Testing (or 'hall testing' as it is also known) environment may approximate normal consumption and hence could be considered.

Note 1: The use of central locations / halls may be appropriate for exploratory/early product development research and will be covered in more detail in a separate guideline. However, if hall testing is used in these early stages, it **must** be followed by a normal in-use test.

Note 2: If the purpose of a test is to examine some special characteristics of a food product that requires strict controls (e.g. a very precise cooking method), then a central location/hall test may be appropriate. Such tests must be followed up by a final in-home product test.

Note 3: Central location/hall testing is not normally appropriate for HPC products, even for early stage testing because:

- the product is tested under unreal conditions;
- consumers will not get the same experience of the product as they would in normal use;
- short-term, immediate impact sensory characteristics are likely to dominate responses;
- consumers are asked to make an immediate judgement after one experience.

13. Test Length

The length of the test period should be set to ensure the respondent has sufficient time to assess the product adequately. This would involve usage over multiple occasions. The specific test length will depend on frequency of usage of the product field:

- **The allocated time should take in to consideration the number of desired uses to see the benefit. Each category should set its own rules on a country to country basis as appropriate for the product concerned.**
- **For products that are consumed fully on a single occasion (e.g., some foods), single packs should be provided but adequate numbers for multiple usages.**
- **For products with a 'carry over' effect that occurs over several weeks, the length of the test will need to be extended.**

The test length should normally be set in units of a week to ensure adequate representation of each day of the week.

If the product has a 'long term' effect that can only be meaningfully understood by the consumer after repeated use (as in the case of a colour protection benefit provided by a detergent or therapeutic effects of a shampoo), then the test should be long enough so that these longer term benefits have a chance to manifest themselves.

The length of the test should be long enough for the respondent to use the entire sample provided. This is important as sometimes issues related to dispensing from the packaging come up when the product is nearing completion.

14. Test Product

Rigorous quality standards must be applied to production of test product. Ideally test products should be sourced from main plant trials (rather than pilot plants) to ensure they are representative of what will be launched into the market. The test product and the control product should be sourced in a similar manner to ensure that they are comparable.

The only exception to this is a Competitor Benchmarking test where products are sourced from the marketplace.

Input should be received from Development/CTI on the sensory and performance properties of test products to ensure they are a) representative and b) not subject to too much variability. Safety clearance and patent clearance procedures must be carefully followed.

14.1 Quality Standards of Test Product

It is critical that the products being tested are of the correct quality. The quality standards applied to experimental test products should be tighter than those applied to normal production, in order that the effects being tested are not masked by variation within the test samples.

Generally when testing alternative formulations of one of our current or new brands, products should be made from the same batches of raw material and the entire quantity needed for a single test should be produced in a single run. Every effort should be made to ensure that test product is made in the main plant so that it is representative of what will be produced and launched into the marketplace.

Current product, when used as a control, should be taken off the line and thoroughly checked to ensure that the key parameters are within specification and that the variability is no higher than in normal commercial production.

After manufacture, product may need to be aged to simulate the time it spends in the supply chain between the factory, the supermarket shelf, and the consumer’s home. This is particularly important for products like detergents and toilet soap whose physical properties change with ageing. Detergents, fabric conditioners, hair conditioners, dishwashing liquids and household cleaners should be aged for 3 weeks. For other HPC and for Foods categories, technical/development specialists should be consulted to determine the optimum ageing period between production and use in a product test.

For **competitor benchmarking tests**, all products in the test should be representative of what is available in the marketplace. Thus it will be necessary to purchase both competitive and Unilever products from a representative sample of retail outlets consistent with the national distribution profile of the products. Batch numbers should be recorded.

Whilst, products purchased from a single supermarket or wholesaler will give product with low performance variability, they will be unrepresentative of normal production.

NOTE: For Competitor Benchmarking, for renovated products, we can use main plant trial samples, aged appropriately. These products will need to be retested with proper in-market samples as soon as the product is established in market.

14.2 Product Appraisal and Sensory Guidance

Before completion of the Market Research Proposal for a product test, the CMI manager should obtain from the appropriate source (normally development/CTI), a detailed Product Appraisal. This appraisal will provide technical advice on the specific performance properties of the test (and control) products as well as insights on sensory signals that could be important in directing analytical work and fully understanding final results of the test.

14.3 Patent Clearance

If a patent application has not been filed, and in the absence of confidentiality agreements or under circumstances where confidentiality is not implicit, the revelation of a product (or concept) to consumers, anywhere in the world, **renders that product <u>not new</u> under most patent laws, rendering the product in question no longer patentable.**

1.1 *This is generally not an issue in product testing because patent issues will normally have been sorted out before this, but it may be an issue in early stage product testing. In such cases it is the responsibility of Marketing and Development to decide whether a patent issue exists. In such cases, a confidentiality agreement should be signed by all consumers participating in the test – see Appendix 3 for confidentiality agreement.*

14.4 Safety / Microbiological Clearance

For products with a limited shelf life (e.g., foods) or with high water contents (e.g., shampoos, face creams) microbiological clearance is required to ensure the product is safe to consume/use.

It is not necessary to obtain safety clearance on products purchased from retailers unless they have been repacked, provided good care has been taken in their acquisition and distribution.

ESOMAR rules are to be followed for ethical guidelines

15. Product Packs and Quantities

The amount of product placed with consumers must be sufficient to meet their needs over the placement period. This will normally be a single pack of the most widely used size.

For most product tests, you should allow an extra 20% for over-recruitment of consumers. For online studies a higher percentage will be needed.

Marketing and R&D are responsible for providing test product stimulus.

In some product categories, especially Foods, special transportation arrangements need to be made to ensure that the product quality is maintained. Where relevant, guidelines will be provided by the Category teams.

In some countries the law requires other information such as an ingredients panel to be provided in all tests, including unbranded testing. If so, this must be clearly shown. Detailed usage and cooking instructions should be given when appropriate.

When re-packing Unilever or competitive product, product code, use-by date, batch number, and source of purchase should all be coded onto the new pack.

At placement, consumers should be instructed that if they experience any adverse reaction to the product to seek medical advice immediately. Where required by law, respondents should be given a telephone number to contact if they experience any adverse reaction to the product under test. This is a rare but extremely serious event, and appropriate procedures should be established within the company to deal with it.

NOTE: Any adverse reactions caused as a result of products under test are a PR risk and should be dealt with appropriately.

16. Mode of Interviewing

The interviewing mode utilised is determined once a test design has been determined and is completely dependent on the objectives of the study. The key questions to answer include:

- **Is there a need for an interviewer administered questionnaire, or is self-administered acceptable?**
- **Is the target to be interviewed more accessible or more likely to respond to personal, mail/postal or internet interviews?**
- **In a comparative test, will receiving both products at the same time bias results in some way?**
- **Does the unused product need to be returned for security reasons?**

A variety of interviewing approaches may be utilised once test design has been determined. The following are some examples:

- Door-to-Door product placement followed by personal interviews;
- Test Centre / Mall placement with personal or telephone interviews;
- Panelists placed with product via post/mail with self-administered interviews on PC or mobile device.

Interviewing approaches will depend upon the country, test objectives and sample design (e.g. low penetration target sample groups).

If there are issues of literacy or clarity of test procedure, then personal interviewing rather than self-administered postal questionnaires should be used.

In tests where consumer memory or recall is likely to be difficult, it can be helpful to use **leave-behind questionnaires** or “feed-back” sheets.

Internet and postal methods do work for product testing, particularly among consumer panelists recruited for general commercial research purposes. In fact there are good reasons to employ them as there is growing evidence that high quality testing can be achieved using self-completion questionnaires at relatively low cost. However, care must be taken to only work with reputable, well-managed panel operations and where panel members have not been over-exposed to surveys in our category of interest (a gap of at least six weeks between tests is desirable). When recruiting minority samples, discussions must be had with the research agency to establish feasibility vis a vis panel size, panel usage and specific recruitment criteria.

17. Questionnaire

There is a standard questionnaire format and questioning sequence that MUST be used. (See Appendices 1-2).

Categories will need to customise the standard questionnaire to take account of specific functional attributes that are known to be critical to product acceptance and long-term brand choice in the Category.

Globally harmonised lists of core functional product attributes should exist for all Categories and Managers should contact their Global/Regional Category Leader to find out what these are. Unilever teams should ensure those lists are updated in their category questionnaire held with the suppliers for Product Testing.

While Managers should feel free to add attributes beyond this core grouping (to address specific demands of the test), the total number of attributes should be kept as short as possible and should certainly not exceed 30.

For each of the basic methods of product testing, there is a standard sequence of activities and questioning required to collect information on reaction to the product, either singly or in comparison to another product.

The recommended design sequences are as follows:

Monadic Tests	Paired Comparison for VIP	Sequential Monadic
Screening	Screening	Screening
Product Placement	Product Placement	Product Placement
Product Usage	Product Usage (1)	Product Usage (1)
Questioning	Questioning (overall opinion only)	Questioning
	Product Placement	Product Placement
	Product Usage (2)	Product Usage (2)
	Questioning (overall opinion Only)	Questioning
	Comparative Questioning	Comparative Questioning

17.1 Questionnaire length

The length of the questionnaire (screener / concept + recall), inclusive of profiling questions (e.g LSM), should be no longer than 50 minutes, and the recall portion of the questionnaire cannot be longer than 25 minutes. Questionnaires with a longer duration dramatically reduce the quality of the responses obtained from the respondents.

Each category must define a set of 15 core attributes per product that will be used on all tests. This will also enable data mining over a period of time. For instance, the Oral care team should define a core set of attributes for the family health brand whitening product, family health brand core product etc.

The questionnaire should consist of 15 core attributes and up to 15 project specific attributes.

18. Questionnaire Design

For each test method the questionnaire structure is essentially fixed and has specific wording of questions, with only minor modifications to the questionnaire as required for the particular product being tested. An explanation of the questionnaire structure and a specimen questionnaire is given below for:

- **Monadic tests**
- **Paired comparison (for VIP tests)**
- **Sequential monadic tests (benchmarking)**
- **Concept – product tests**

Reference is made below to the attributes to be used in product testing
Please note that these relate to a limited set of key overall attributes that must be used in all product tests. They do not cover category specific attributes.

18.1 Monadic Product Test

18.1.1 Questionnaire Overview

Questions MUST be asked in this order

Confirm Usage of test product	
Overall Opinion	(7 point scale)
Likes	(open ended)
Dislikes	(open ended)
Value for Money (if priced)	(5 point scale)
Uniqueness (if relevant – i.e. a new brand or major change to an existing brand)	
Comparison with expectations	
(if concept has been shown)	(5 point scale)
Attribute ratings	

- Key overall uni-polar attributes
- Other uni-polar attributes
- Bipolar attributes
- Specific Diagnostics

18.1.2 Questionnaire Design Issues

The wording to be used for each topic and the scales to be used are shown in the outline questionnaire in Appendix 1. The following points should be noted in particular:

Uniqueness
This should only be asked for a new product/brand or a major change to an existing brand.

Scale direction

For all countries outside North America the scales must run **negative to positive**, to avoid ‘top-boxing’ phenomena and to maintain test sensitivity.
For North America only certain scales(see Appendices 1-3) will be ***reversed*** i.e. run from positive to negative because…

1) it is believed that NA consumers intuitively process information in that direction
2) to maintain marketing familiarity with previous results

Note that this requires two separate questionnaires for cross regional tests involving North America

Attribute Ratings

There are two main types of attributes

1 Uni-polar
2 Bi-polar

1. Uni-polar: Those where the more the product has the attribute the better it is (e.g.,, quality, good taste, removes stains).

2. Bi-polar: Those where no particular level of the attribute is the correct level for all respondents (e.g., degree of sweetness, amount of lather).

Whenever the attribute is not obviously uni-polar, bipolar attributes should be used as they give more guidance on what needs to be done to improve the product (e.g., use a bipolar ‘too sweet…. not sweet enough’ rather than a uni-polar ‘has the right level of sweetness’).

Uni-polar attributes can further be split into two types:

Key dimensions overall
Other uni-polar attributes

Key dimensions overall ***are measured using a*** *7-point overall opinion scale* ***modified to refer to the attribute. The dimensions to which this scale is applied are limited:***

Foods	*HPC*
Quality	Quality
Taste	(Fragrance)
Appearance	

Other uni-polar attributes are measured **using a 5-point agree-disagree scale.**

For **bipolar attributes** using **a 5-point scale** are constructed so that the mid point of the scale represents the ideal or most desirable response and either side of the mid-point indicates too much or too little of the attribute. For bipolar attributes the precise wording will depend on the specific attribute, for example:

"Thinking about xxx, would you say it is…?"

Much too	(strong)
A little too	(strong)
Just right	
A little too	(weak)
Much too	(weak)

Much too much	(lather)
A little too much	(lather)
Just enough	
Not quite enough	(lather)
Not nearly enough	(lather)

18.1.3 Outline Questionnaire for Monadic Tests

See Appendix 1.

18.2 Paired Comparison for VIP

18.2.1 Questionnaire Overview

Questions MUST be asked in this order

The Paired Comparison for VIP questionnaire is specifically for final validation of a VIP product

First Recall

Confirm Usage of test product
Overall Opinion (7- point scale)

The second product is then placed

Second Recall

Confirm Usage of test product
Overall Opinion (7 - point scale)
Overall Preference (3 - point scale)
Preference on key attributes (3 - point scale)

18.2.2 Questionnaire Design Issues

The questionnaire is designed to be succinct and to provide final sign off for a cost reduced/harmonised/flexible product. This questionnaire is not designed to provide lengthy diagnostics for product improvement

18.2.3 Outline Questionnaire

See Appendix 2

18.3 Concept-Product Test

18.3.1 Overview

The Concept Product Test methodology is used primarily at the end of the Feasibility phase of the Innovation Funnel instead of a full Simulated Test Market for low investment projects.

18.3.2 Concept-Product Test Methodology

The Concept Product test is conducted branded and monadically. Respondents are shown the concept on a concept board and asked questions based about it. Concept Product tests should be conducted among top 3 box acceptors on the purchase intent scale. They are then given the product to use and are asked further questions in a recall interview.

There are differences between the recommended STM methods in how the price of the brand is introduced. The relevant STM approach for handling price should be used in the Concept-Product Test. For Global projects one approach (priced or un-priced) must be agreed in advance.

A priced concept is preferred. If a final decision is being made with no-follow-up STM the concept must be priced. The concept can be un-priced if it is not a final test.

The questions are generally the same at the concept and recall interview so that results can be compared between the two stages. This helps understand whether reactions to the product match, exceed or fail to reach those generated by the concept.

18.3.3. Outline Questionnaire for Concept-Product Test

Questions MUST be asked in this order

The specific interview flow for a concept-product test will typically cover the following:

CONCEPT INTERVIEW

Quota and screening questions
Product usage questions
Overall opinion
Purchase interest
CONTINUE ONLY WITH TOP 3 BOX ACCEPTORS AT PURCHASE INTEREST
Likes/dislikes
SHOW PRICE IF NOT ALREADY GIVEN
Purchase interest (if price not already given)

Value for money

Uniqueness
Expected frequency of purchase
Attribute battery
GIVE PRODUCT TO TAKE HOME

RECALL INTERVIEW

Did you use the product? If so, how much? If not, why not?
Overall opinion
Likes/dislikes
Value for money
Uniqueness
Comparison with expectations
Expected frequency of purchase
Attribute battery
Specific diagnostic questions

19. Drivers Analysis

We can identify 2 types of drivers: category product drivers and individual product drivers. The two types of drivers are different and complement each other.

Category Product drivers also referred to as “macro level analysis”: These are drivers of liking that operate over a range of brands / products in a category. Category Drivers Analysis is conducted on multiple products (usually 4+), whether in a single test or multiple tests. The objective of this analysis is finding over-arching elements that drive liking in the selected range of products and guiding development towards an optimised product. A representative set of brands is needed to reference this analysis as category drivers.

Individual Product drivers also referred to as “micro level analysis”: These are determined from consumer responses to a single product and rely on their diagnostic attribute ratings of which attributes need to be changed and in what direction. This is typically done in the context of understanding / diagnosing issues with a particular product / formulation.

The following are the only drivers analysis that should be used for product tests

Macro level analysis –

- Factor analysis followed by regression
- Biplots

Micro level analysis –

- Relative Weight Analysis
- Penalty analysis
- Difference Drivers

19.1 Factor analysis followed by regression

In a product test where we have a set of attributes evaluated by consumers and the aim is to identify which are drivers of product liking/opinion, it is very likely that there will be strong inter-relationships between these attributes. Via Factor analysis attributes are combined into a number of components which are now statistically independent. The components are modified, by a process referred to as rotation (the varimax method which keeps statistical independency between factors or components), to allow similar attributes to associate together with specific components. The result provides a clearer picture of attribute association. These components may then be used as input to regression analysis subsequently, to assess the relationship with overall opinion.

19.2 Biplots

A biplot is a macro level analysis (mathematically very similar to principal components analysis) that is a form of a perceptual map. It is a way of representing the products and attributes in a two-dimensional space.

Biplots provide visual summary of information on:

- Differences and similarities among objects – Products that are closer to each other are more similar as compared to products that are farther from each other.
- Attributes that best discriminate between products – The length of the vector of the attribute indicates the degree of discrimination. The longer the vector the more discriminating the attribute is.
- Attributes that are similar to each other – The angle between the vectors of the attributes indicates similarity between attributes. The smaller the angle the greater the degree of similarity. An angle of 90 degrees indicates the attributes are not correlated and an angle of 180 degrees indicates the attributes are negatively correlated.
- Relative rating of products on attributes – The relative rating of products on an attribute is obtained by projection. For the attribute of interest, extend the vector beyond the origin and draw perpendiculars from the products onto the attribute vector. The distance between the origin and the point at which the perpendicular intersects the vector represents the relative performance/rating of the product on the attribute.
- The greater the distance with the similar direction with the overall liking vector, the higher is its performance/rating on the attribute.

19.3 Penalty Analysis

Penalty Analysis is a micro level analysis which shows the impact of failing to achieve "the right level" of a specific attribute.

Penalty analysis has two components that contribute to interpretation:
1. For each product characteristic evaluated using a "just right" scale, the first component is the percentage distribution of responses.
Respondents may view the product characteristic as "just right," as having "too little" of the characteristic, or having "too much" of the characteristic.

Penalty analysis requires that the survey uses "just about right" (JAR) 5 points scales with percentages:

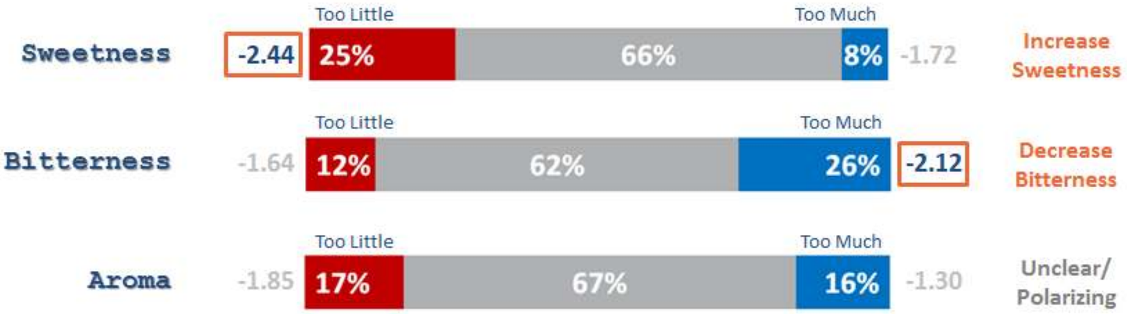
Just about right (JAR)

Much too strong	+2
Slightly too strong	+1
Just about right	0
Slightly too weak	-1
Much too weak	-2

However, these scales are collapsed to 3 points to accommodate stable analysis and interpretation.

2. The second component is the penalty itself. A penalty is defined as the decline in a dependent measure (e.g. Overall Opinion) between those respondents who specify the product as "just right" and those who said the product possesses "too little" or "too much". As such, two penalties are estimated. One for each of the directions away from "just right".

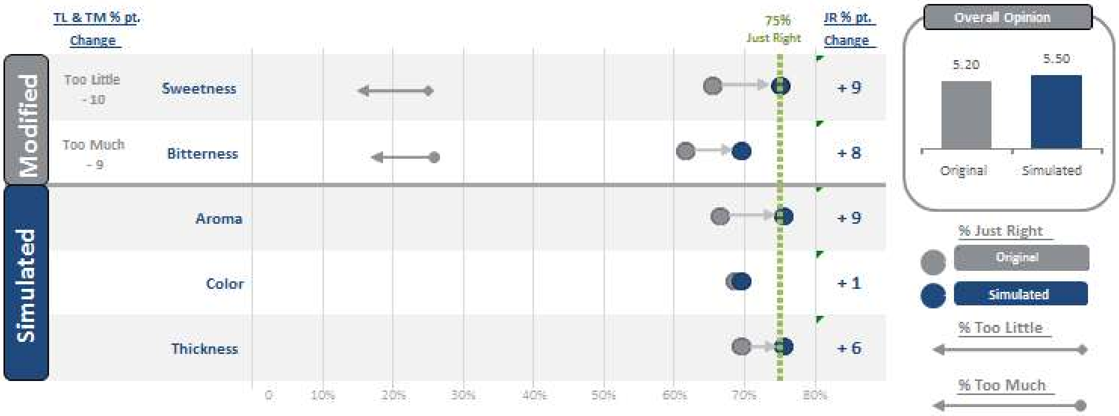
Example of penalty analysis:



Therefore the analysis determines the “penalty” a product pays for not being perceived as “just right” on a particular characteristic/attribute. This identifies product attributes which, if modified, could improve Overall Opinion. The change in Overall Opinion is displayed based on the simulated impact of these modifications.

In the above example, the product is found to be not sweet enough (i.e. the level of “sweetness” is too low / too little), and therefore the recommendation would be to increase this. For “Bitterness” it is the other way around, and the level of bitterness needs to be decreased.

This modification scenario (increase sweetness and decrease bitterness) is input for the simulator (refer to the grey arrows):



To understand just how much adjustment is needed, a balance must be found between an adjustment that would be too large and would alter the product profile too much, and one too small that would not impact overall opinion.

In this case a 10 %-point increase in sweetness (which means the product profile is altered in such a way that 10% of respondents who regarded the product as ‘too sweet’ would now rate the sweetness to be ‘just right’) and a 9 %-point decrease in bitterness were found to be ideal.

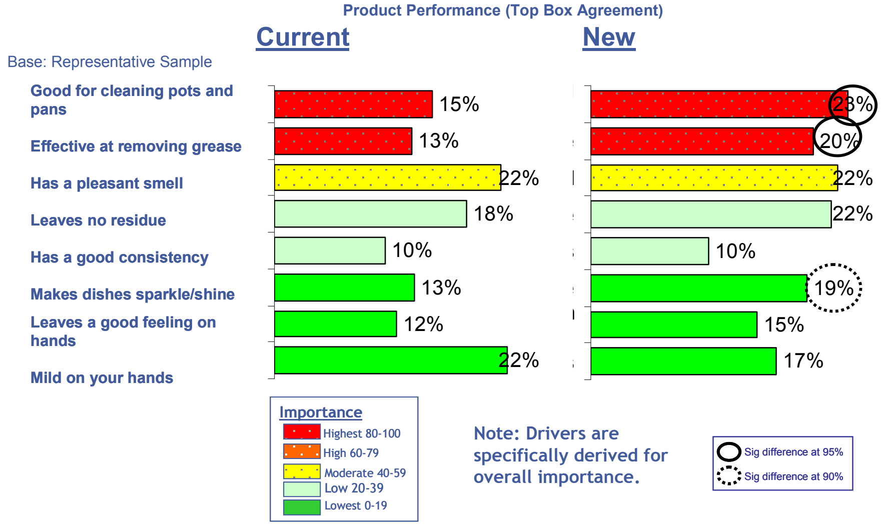
This profile change results in a 9 %-point increase in just right for sweetness (just 1% of the sample who before the change regarded the product to be just right now regards the product ‘too sweet’) and an 8 %-point increase in just right for bitterness.

Subsequently these changes also increase the just-right perception of other product characteristics, most notably aroma and thickness, and ultimately overall opinion, which mean score increases from 5.20 before to 5.50 after the change in profile.

19.4 Relative Weight Analysis (RWA)

RWA is a form of regression analysis that attempts to provide meaningful weights to attributes. RWA uses a singular value decomposition approach to transforming the original set of correlated attributes into an equal number of uncorrelated transformed variables that retain as much similarity to the original variables as possible. The regression to Overall liking is conducted on these transformed uncorrelated variables and the results are expressed in terms of the original attributes through a reverse transformation process.

In the example below the RWA Relative importance is used to color code the performance of the Current and New products. By understanding which Attributes are most strongly related to Overall Opinion along with the relative performance key strengths and weakness of the products can be understood.



*This chart is used for illustrative purposes only and contains blinded data

19.5 Difference Driver Analysis and T-plot

The goal of the difference driver analysis is to identify a small set of attributes that have the greatest impact on “driving” the difference between two products in overall opinion. Combined with the T-plot it provides a visual snapshot of this impact.

An attribute is considered impactful in “driving” the difference in overall opinion ratings between products if, when the attribute ratings are statistically held constant, meaning the attribute ratings are brought to the same level for both products, the difference in overall opinion ratings changes the most.

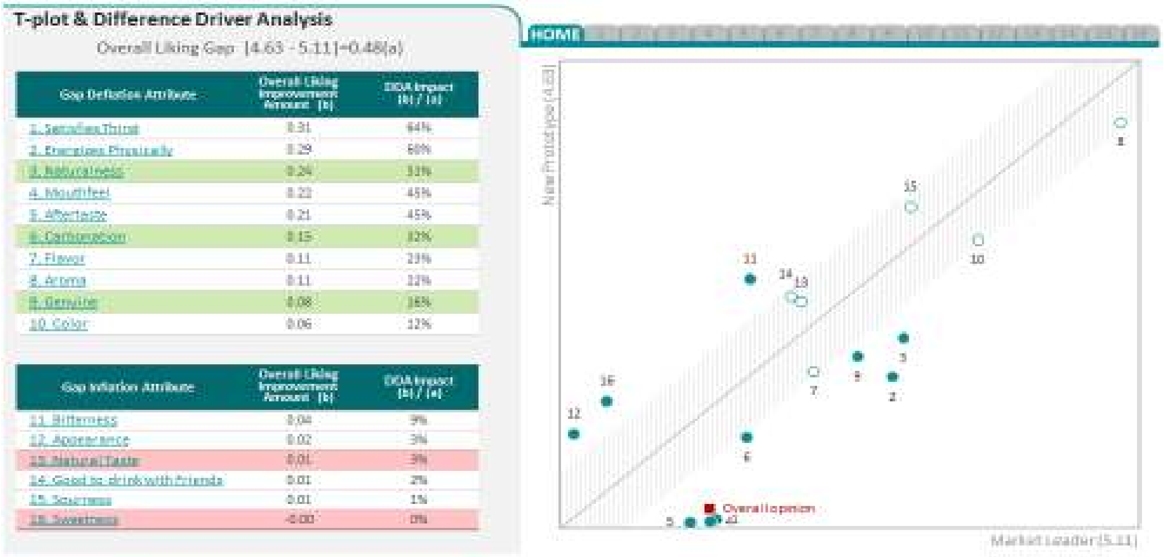
For example, consider two products having significantly different average ratings on both overall fragrance opinion as well as overall opinion. Overall fragrance opinion will then be considered an important attribute, if when fragrance ratings of both products are statistically adjusted to be equal then results in a large reduction of the difference in overall opinion averages. It was the difference in fragrance then that drove the difference in overall opinion.

The analysis is performed separately for each attribute. For each attribute, the analysis takes into account:

- The percentage point or mean difference between attribute ratings for the two objects.
- The percentage point or mean difference between overall acceptance ratings.
- The correlation the attribute has with the overall acceptance measure.
- Then assesses the impact the attribute has on the difference in overall acceptance ratings (either reduces or increases the difference).

A measure of change is calculated such that the attributes which are more impactful in “driving” differences in the overall opinion ratings are assigned a greater weight (or value). The weight (or value) for each attribute is interpreted as an impact score. The weight (or value) assigned to each attribute is a percentage change in the difference between products in overall opinion ratings.

- A gap deflation attribute “drives” a reduction in the difference in overall acceptance ratings (makes the difference smaller).
- A gap inflation attribute “drives” an Increase in the difference in overall acceptance ratings (makes the difference larger).



T-plot & Difference Driver Analysis

Overall Liking Gap: (4.63 - 5.11)=-0.48(a)

Gap Deflation Attribute	Overall Liking Improvement Amount (b)	DDA Impact (b) / (a)
1. Satisfies Thirst	0.31	64%
2. Energizes Physically	0.29	60%
3. Naturalness	0.24	51%
4. Mouthfeel	0.22	45%
5. Aftertaste	0.21	45%
6. Carbonation	0.15	32%
7. Flavor	0.11	23%
8. Aroma	0.11	22%
9. Genuine	0.08	16%
10. Color	0.06	12%

Gap Inflation Attribute	Overall Liking Improvement Amount (b)	DDA Impact (b) / (a)
11. Bitterness	0.04	9%
12. Appearance	0.02	3%
13. Natural Taste	0.01	3%
14. Good to drink with friends	0.01	2%
15. Sourness	0.01	1%
16. Sweetness	-0.00	0%

19.6 Optional analyses

When to use a single driver analysis

Under certain circumstances it can make sense only to do some of the analyses outlined above, such as RWA, if we have only 2 cells or there is not a lot of variability in the data.

Appendices

Appendix 1 - OUTLINE QUESTIONNAIRE FOR MONADIC PRODUCT TEST

RECRUITMENT INTERVIEW

These questions should be modified as required for the particular product field being tested.

Security Screening

1. Today we are looking for people who work in certain types of job. Do you or any member of your family work in any of these areas?

   SHOW CARD

- Advertising
- Product Manufacture of (insert category) products
- New Product Development
- Design Agency
- Perfume House
- Construction
- Military
- None of these

2. Do you or any members of your family work for any of these companies or organisations?

   SHOW CARD which lists specific competitive or sensitive organisations in these areas:

   List a total of 3-5 major competitors, industry suppliers or retail groups

   None of these

   CLOSE INTERVIEW if any competitive or sensitive companies are mentioned.

Recent Test Involvement

3. When was the last time you participated in a survey (test) like this?

   CLOSE INTERVIEW if more recent than three months *(internet will have a shorter time frame)*

Category Usership

4. "Which of these products do you use regularly?"

   SHOW CARD listing test product category and related categories

   CLOSE INTERVIEW if test product category not mentioned

Heaviness of Use

Category-specific questions may be inserted here to determine the respondent's heaviness of use, e.g., Number of washloads per week, frequency of shampooing hair, eating ice cream, etc.

Responses may be used for purposes such as screening out abnormally heavy or light users or balancing cell-to-cell representation of user types.

Some questions to assess the environment in which the product is used will be useful, where it impacts product performance (i.e. Water hardness levels etc)
Also based on category usage, the habit questions like Wash habits (i.e. Use of Bar & powder or powder alone etc., use of any cleaning/body cleansing implements etc). Some of these may become key panel matching criteria.

Brand Usership

4a. AMEND AS APPROPRIATE TO STUDY Which of these brands of (test product category) have you ever bought?

4b. AMEND AS APPROPRIATE TO STUDY Which of these brands of (test product category) have you purchased in the past six months?

5a. Which of these brands do you use most often?

5b. Which of these brands would you never consider using?

5c. Brand Loyalty

Example: "Try to imagine the last 10 times you purchased (test product category). For the brands listed on this card, please tell me the number of times you purchased each brand over the last ten purchase occasions by giving it a zero if you do not recall buying the brand at all or a number from 1 to 10 if you bought it once or more. For instance, if you remember buying two brands with about equal frequency, give each 5 points. If you are certain you only purchased one brand over the past 10 times give that brand 10. The numbers you give these brands must total 10".

Household Characteristics

Questions may be inserted here related to household characteristics such as household size, income, or age. Responses may be used to screen from the sample respondents with characteristics not desired in the test sample (e.g.,, single people not being appropriate targets for the test brand) or responses may be used to assure reasonable representativeness of such sub-groups in a multiple cell test.

Request for Co-operation

For those individuals qualifying for the home use test based upon these kinds of screening criteria, a question should be asked at this juncture requesting cooperation in a test, involving usage of a product in the home and a follow-up interview. Duration of the test should be clearly stated and any incentive offered should be fully explained.

Placement Instructions

Co-operating respondents should, at this juncture, be given a set of verbal instructions about the basic "mechanics" of the test in which they have agreed to participate. This should include information about the duration of the test and call-back interviews.

Confidentiality

If required by conditions of extraordinary security, it would be at this point when the respondent would be asked to agree to any confidentiality or product return arrangements

RECALL INTERVIEW

Confirming Product Use

1a. (For Multi-Use products) First of all, can you tell me how much of (the test product) you have used since you received it? **OR**

(For Single-Occasion use products) First of all, can you tell me whether you have (used/eaten) the (test product) since you received it?

IF 'NONE' ASK Q.1b

1b. What is the reason you have not used/eaten any of the (test product)?

RESCHEDULE RECALL INTERVIEW OR CLOSE INTERVIEW IF RESPONDENT IS UNWILLING TO USE/EAT TEST PRODUCT

Overall Opinion

2 Taking everything into consideration, which of these phrases best describes your overall opinion the (brand/product) you have just (tried/used)?

SHOW CARD FOR SCALE: REVERSE SCALE FOR N AMERICA
Very Poor
Poor
Neither Poor nor Fair
Fair
Good
Very Good
Excellent

Open-Ended Likes

3. What, if anything, did you particularly like about (test product)?
PROBE UNCLEAR OR AMBIGUOUS ANSWERS
Why do you say that?
PROMPT “What else?” TWICE (MAXIMUM)

Open-Ended Dislikes

4. What, if anything, did you dislike about (test product)?
PROBE UNCLEAR OR AMBIGUOUS ANSWERS
Why do you say that?
PROMPT “What else?” TWICE (MAXIMUM)

Value for money (if priced product)

5. “Considering this price of xx for (INSERT QUANTITY/SIZE), which of these phrases best describes how you feel about the value for money of this (brand/product name)?”

SHOW CARD – UNIQUENESS - REVERSE SCALE FOR N AMERICA

Very poor value for money
Fairly poor value for money
Average value for money
Fairly good value for money
Very good value for money

Uniqueness (if relevant to include)

6a. Thinking about the uniqueness of this (brand/product name), which of these phrases best describes your opinion?

SHOW CARD – UNIQUENESS - REVERSE SCALE FOR N AMERICA

Same scale as BASES uses

It is not at all new and different
It is slightly new and different
It is somewhat new and different
It is very new and difference
It is extremely new and different

Comparison with expectations (if relevant to include)

6b. Thinking about how this (brand/product name) compared with your expectations, which of these phrases best describes your opinion?

SHOW CARD - REVERSE SCALE FOR N AMERICA

It is much worse than expected
A little worse than I expected
As I expected
A little better than I expected
Much better than I expected

Frequency of use (if relevant to include)

6c. How often, if ever, do you think you would buy this (brand/product name)?

AMEND AS APPROPRIATE FOR PRODUCT AREAS

Every day or more often
4-6 times a week
2-3 times a week
Once a week
Once every 2-3 weeks
Once a month/every 4 weeks
Once every 2-3 months
Once every 4-5 months
Once or twice a year
Less often than once a year
Never

Attribute Ratings

7a. **Overall Uni-Polar Attributes**
(“Quality”, “Taste”, “Appearance” for Foods)
(“Quality”, “Fragrance” for HPC)

I would like to have your opinion of various characteristics of the (brand/product name) you have tried. For each characteristic I would like you to tell me which phrase (on this card) best describes your opinion of (brand/product name). There is no right or wrong answer. It is your personal opinion we are interested in.

READ CHARACTERISTICS IN ORDER BELOW

REVERSE SCALE FOR N AMERICA

Very Poor
Poor
Neither Poor nor Fair
Fair
Good
Very Good
Excellent

7b. **Other Uni-Polar Attributes** (These attributes should be more specific than the other Overall Uni-Polar Attributes above and should represent all themes)

Now I am going to mention some (other) characteristics that might be used to describe the (brand/product name) you have tried. As I mention each characteristic, please tell me which phrase (on this card) best describes your opinion of the (brand/product name).

FOODS - READ CHARACTERISTICS ONE BY ONE IN ORDER BELOW
ORDER TO FOLLOW LOGICAL FLOW
HPC – READ CHARACTERISTICS ONE BY ONE STARTING WITH THE X’ED CHARACTERISTIC – *RANDOMISED ORDER*

SHOW CARD – AGREE/DISAGREE SCALE

Disagree Strongly
Disagree a Little
Neither Agree Nor Disagree
Agree a Little
Agree Strongly

7c. **Bi-Polar Attributes**

“Now I am going to mention some (other) characteristics that might be used to describe the (brand/product name) you have tried. As I mention each characteristic, please tell me which phrase (on this card) best describes your opinion of the (brand/product name).

SHOW CARD – MID-POINT OPTIMAL SCALE

EXAMPLES:
(“Strong Flavour…Weak Flavour”)

Much too	(strong)
A little too	(strong)
Just right	
A little too	(weak)
Much too	(weak)

OR (e.g.,, “Amount of lather…..”)

Much too much	(lather)
A little too much	(lather)
Just enough	(lather)
Not quite enough	(lather)
Not nearly enough	(lather)

Diagnostic Questions

Optional questions may be added at this point in the questionnaire to address issues specific to the product assessment at hand. Such special diagnostic questioning should not precede any of the mandatory questioning shown above.

BRAND RECOGNITION QUESTIONS <u>ONLY FOR BLIND/UNBRANDED</u> TESTS

1. Did you recognise the product we gave you to try?
No, I did <u>not</u> recognise the product GO TO NEXT SECTION
Yes, I did recognise the product GO TO NEXT QUESTION

2. What did you think it was?
DO NOT READ NOR SHOW LIST
Precoded list as appropriate to study

Appendix 2 - OUTLINE QUESTIONNAIRE FOR PAIRED COMPARISON FOR VIP TEST

SHOW CARD - use responses to quota for certain brand usership sub-groups or to screen out brand rejectors

RECRUITMENT INTERVIEW

These questions should be modified as required for the particular product field being tested.

Security Screening

1. Today we are looking for people who work in certain types of job. Do you or any member of your family work in any of these areas?

   SHOW CARD

- Advertising
- Product Manufacture of (insert category) products
- New Product Development
- Design Agency
- Perfume House
- Construction
- Military
- None of these

2. Do you or any members of your family work for any of these companies or organisations?

   SHOW CARD which lists specific competitive or sensitive organisations in these areas:

   List a total of 3-5 major competitors, industry suppliers or retail groups

   None of these

   CLOSE INTERVIEW if any competitive or sensitive companies are mentioned.

FIRST RECALL INTERVIEW

Confirming Product Use

1a. (For Multi-Use products) First of all, can you tell me how much of (the test product) you have used since you received it?

OR

(For Single-Occasion use products) First of all, can you tell me whether you have (used/eaten) the (test product) since you received it?

Overall Opinion

2 Taking everything into consideration, which of these phrases best describes your overall opinion the (brand/product) you have just (tried/used)?

SHOW CARD FOR SCALE: REVERSE SCALE FOR N AMERICA

Very Poor
Poor
Neither Poor nor Fair
Fair
Good
Very Good
Excellent

PLACE SECOND PRODUCT

Second Recall

Conduct second recall interview; repeating Overall Opinion (for the second product)

- Then ask preference questions

1. Taking everything into consideration, which of the two products you tested did you prefer overall?

SHOW CARD

I preferred the product I tried first
I preferred the product I tried second
I have no preference

Preference on Attributes

TO PERMIT COMPARISONS AND ESTABLISH PREFERENCE FOR THE PRODUCTS ON SPECIFIC ATTRIBUTES, ASK

3. Even though you may have preferred one of these two products overall, you might prefer one or the other for specific characteristics. I am going to read you a list of specific characteristics that might be used to describe (category/product). Of the two products you tried, I would like you to tell me which you prefer for that characteristic. Which do you prefer for...

READ “X’ed” CHARACTERISTIC FROM LIST or SHOW CARD, DEPENDING UPON RE-CONTACT METHOD

I preferred the product I tried first
I preferred the product I tried second
I have no preference

CONTINUE READING LIST, WITH “WHICH DO YOU PREFER FOR” PREAMBLE UNTIL COMPLETED

Appendix 3 - CONFIDENTIALITY AGREEMENT

A. SAMPLE INDIVIDUAL CONFIDENTIALITY AGREEMENT

(Sample document before using please ensure this is in adherence with your local laws)

PRODUCT TEST CONFIDENTIALITY AGREEMENT FOR RESPONDENTS

Thank you for your interest in participating in this research. It is important that you understand the conditions under which you will be agreeing to participate. Please read the following carefully, as you will have to confirm you have done so and agree to these conditions before you can participate in this product test. You should also keep it somewhere safe so that you can refer back to it during the study. If you have questions regarding any aspect of this research study, or the following conditions of participation, please discuss them with the interviewer before agreeing to take part.

Who is carrying out this study?

This research is being carried out by [insert name of Ipsos entity] ("Ipsos") on behalf of [INSERT EITHER: the client name OR "our client that makes and/or distributes the product"] (the "Sponsor").

What do I have to agree to?

By participating in this research, you understand and agree that you are consenting to the following terms and conditions of participation. For any study where we invite you and your family to take part in this product testing, you also understand and agree that you are responsible for making sure each member of the family also has access to, accepts and complies with these conditions; and that you are also consenting on behalf of your family, all of whom have agreed to participate and to you consenting to these terms on their behalf.

1. Participation:

   You confirm that you are of legal age to consent to participate in this research. You agree to read all the instructions provided, and to heed any warning labels or other safety information Ipsos or the Sponsor may provide. You also confirm that neither you, nor any other member of your family that may assist with the product test:

   - use any drugs or medicines; or
   - suffer from any allergies or other health conditions;

   that may make consumption, use of, or contact with the product(s) unhealthy, dangerous or inadvisable for any reason.

   You also agree to personally participate in this research, including personally carrying out any product testing. If for any reason you cannot perform or complete the test, do not get someone else to perform the test for you, instead, please contact us immediately using the contact details we have given to you. If you experience any adverse reaction to the product please seek medical advice immediately. Where required by law, you will have been given a telephone number to contact if you experience any adverse reaction to the product under test.

2. Confidential Information:

   Your responsibility: - Whilst taking part in this study, you may be provided with information about the Sponsors new concepts, packaging, marketing materials and products that has not been advertised, published or marketed. Such information is "Confidential Information" which can only be used for this research. With the exception of any family members taking part in the product testing, you must not:

- Allow anyone else to use the product(s) or see the Confidential Information that has been passed to you,
- Make copies of it, photograph, and video or in any other way reproduce any of the Confidential Information.
- Give the product or other confidential information to anyone else, or publish pictures, video or any other information about the product or other confidential information anywhere, including on social media (e.g. Facebook)

If you breach any of these confidentiality requirements, the harm to Ipsos and the Sponsor would be irreparable. You therefore understand and agree that if you do breach these confidentiality requirements, Ipsos and/or the Sponsor may take action, including action against you. In legal terms, this includes the right to obtain preliminary injunctive relief against any such breach or threat of such breach.

Ipsos's responsibility: - Ipsos will keep the information you provide confidential and will ensure only fully anonymous aggregated data is included in any research findings. Ipsos will not share with the Sponsor your name or any other information that identifies you without your prior explicit consent. Any contact information you provide will only be used for product shipments, to contact you in connection with this and future studies.

3. Ownership:

The concepts, products and packaging shown to you, described to you and/or used by you remain the property of the Sponsor. In addition, any ideas, improvements, discoveries, or inventions that may be generated from this research are also the property of the Sponsor and you agree to assign to Sponsor all ideas, improvements, or inventions resulting from your participation in this research.

4. Waiver, No guarantee or Warranty:

Neither Ipsos nor the Sponsor provide any guarantee or warranty for any product(s) made available for testing. Specifically Ipsos and the Sponsor disclaim any warranty of fitness, merchantability, safety and the like. The product(s) may have only a limited life and other characteristics which may be unknown to us. Therefore, <u>you specifically agree to waive and release all claims against Ipsos and the Sponsor</u> arising out of or related to your participation in this research and your use of the product(s). To the extent permitted by law, neither Ipsos or the Sponsor shall be liable for (i) any injury, death, property damage or other damage sustained or allegedly sustained by you resulting from the distribution, consumption, use of or contact with the product(s) made available in connection with this research, except for and solely to the extent of any injury or damage which may be caused by the Sponsor's or Ipsos' gross negligence; and (ii) your participation in this research.

Appendix 4 - ACTION STANDARDS TEMPLATE

	Action Standard	Current (Ragu Corn Syrup) (A)	Test 1 (Ragu 1% Sugar) (B)	AS	Test 2 (Ragu 1.5% Sugar) (C)	AS
Overall Opinion *mean*	Superior to current @ 95% 2-tailed	5.79	5.90	✗	6.12 A	✓
Sweetness *mean*	Superior to current @ 95% 2-tailed	4.33	4.29	✗	4.40	✗
Freshness *mean*	Pairty to current @ 80% 1-tailed	4.16	4.49 A	✓	4.54 A	✓
Other	Fill-in					

Action Standard Met ✓

Action Standard Unmet ✗

Appendix 5 - POWER PAGE TEMPLATE

REPORT_POWER PAGE_CMI Template_	REPORT_POWER PAGE_CTI Template_

Appendix 6: Master Protocol Research: Cross-Category Benchmarking Program - Star Wars

Category	**[please complete]**
Products (sub category and brand)	- [please complete] - [please complete] - [please complete]
Cross-category benchmarking lead Unilever	[Contact and details]
Cross-category benchmarking lead Research Agency	[Contact and details] to be provided by Research Agency

Instruction on filling in master protocol for your category

Text highlighted in yellow indicates Category specific elements, for each Category to create their Category specific document.

1. Background

Global cross-category benchmarking (also known as competitor or quality benchmarking) is a joint CTI and CMI Initiative.

The scope of the programme for each category is signed off by the GCLT and only entails benchmarking studies that contribute to Unilever's KPIs of product superiority. The results of these studies are collated at a global level and reported to the business in the ROMI Scorecard. **Studies in the program should be labelled as STAR WARS**, to distinguish it from other benchmarking studies (e.g., regional initiatives).

The key determinant of whether consumers continue to choose our products over those of our competitors is the perceived quality and value of those products relative to those to which they are compared. Hence whilst great propositions and promotions can buy us trial, product will have the greatest impact on repeat purchase and loyalty.

Consumers will assess the quality of our products at various points of contact, and all of these will play a role in their propensity to purchase (and repeat purchase). Ideally, any category competitor benchmarking program should encompass all key touch-points to give us a holistic understanding of quality. The quality of our products should be assessed with consumers in store and in home, and also through internal technical or sensory assessment.

When we deliver superior products, it is also important to deliver and communicate this functional superiority in the right way. For this programme, it means that in addition to measuring overall opinion, we also focus on key attributes. We want to dial up the product superiority, which can translate itself into (a) claims and/or (b) benchmarking result improvements.
Key attributes reflect functional attributes that drive the brand experience. Only those attributes that can be translated into a product benefit (either directly claimed or perceived) should be selected.
The key attributes should be part of the BPS and should be differentiating attributes from competition (i.e. Win vs competition). Key attributes can be:

- **Technical: sensorial, clinical, nutritional**

- **Consumer: measured in the blind product test**

The focus of this benchmarking program will be on internal technical assessment and blind testing in an appropriate environment (ideally in-home but where needed Central Location Test).

2. Objectives and Deliverables

The Competitor benchmarking programme has been set up to improve the design quality of our products in the market. It is a snapshot of products in the market, or about to be relaunched in market, to understand how competitive our products are. It is a strategic activity which is different from the development of products within innovation projects. Benchmarking addressing the following questions:

- **How are our products performing against key competitors?**
- **Which products do we need to improve?**
- **Which aspects of our products do we need to improve?**
- **Are we performing according to our Brand Performance Standards (BPS)?**
- **Are we delivering against our key attributes?**

Benchmarking should be viewed as a business scorecard. The test results will tell us how consumers perceive our products without the impact of the brand, relative to the competition and what (if any) improvements are needed to maintain consumer satisfaction and increase sales. It will also allow us to know where and when consumers perceive the benefits we measure internally.

Note on Branded testing

It is the categories discretion whether or not to include branded tests in addition to blind tests. However, only blind product test results will be reported into Compass. In those instances where a blind product test is not possible, the branded results will be reported into Compass.

Branded testing instead of blind testing may only be considered in certain circumstances, subject to the following criteria and is to be approved by Category CMI VP:

Branded testing is only required when:

- It is impossible to remove the brand from the product (e.g. striped toothpaste)
- The brand is in a long term decline and one cannot determine if it is a product or brand issue
- Value Improvement Project (VIP) options need to be explored:
  - Branded test explores VIP opportunities
  - Branded test ensures strengths of the brand before undergoing VIP

3. Scope

The scope of Star Wars benchmarking programme approximates the UEx top 100 cells. The UEx cells represent 56% of NPS, so if we can improve product performance in just these cells, we can already impact a significant part of our business. In selecting scope:

- It is also important to ensure coverage per region

- It does not have to cover top all categories high turnover cells if there is a good reason not to
- It can include products that are of strategic importance / high turn-over outside UEx cells.

The scope of the programme for each category's Star Wars programme is agreed and owned by the GCLT.

4. Briefing

For the briefing, Unilever can use the standard template:

CMI BRIEF - Quantitative Product

CTI BRIEF - Quantitative Product

Unilever should always provide the Research Agency with a complete briefing document.

5. Test Design

The recommended test design for benchmarking is sequential-monadic with a preference question at the end. Unilever calls this approach **"monadic + sequential",** as the monadic ratings (first tried product) are used for assessment against the KPI. The choice for "monadic + sequential" is driven by the ability to get more diagnostic information.

Sequential Monadic
Screening
Product Placement
Product Usage (1)
Questioning
Product Placement
Product Usage (2)
Questioning
Overall Preference Measure = Preference question

The **number of attributes** included during questioning should be maximum 20-25 with at least 10-15 consistent by category.

6. Data Collection

In choosing a data collection method, the following considerations might be useful:

- Costs
- Re-branding/re-packaging issues
- Length of product trial
- Input from an in-home situation/experience, family experience
- Using product in combination with other products

Ideally, the product test should mimic the natural consumer habits and circumstances, and should therefore be conducted in-home. However, in some instances (e.g., contamination issues with repackaging or the need for control in consumer-based evaluation of usage), a Central Location Test may be more appropriate.

A concept or concept statements and other stimuli are not used for competitive benchmarking testing. Using a proposition is allowed (for example "anti-dandruff shampoo", "heart health margarine"), but must then be used for both the Unilever and Competitor product.

7. Consumer Sample and Sample Size

The consumer sample can be either 50/50 (users/non-users) or a general sample of category users (Category users should be defined as narrow as possible in terms of product variant, type, format and consumer need), with boosts of brand users when required (which is a Category decision).
The sample composition is dependent on the Category. CTI and CMI need to reach an agreement on what is best for their Category. When the decision on sample composition is made for a Category, each test within this Category should follow the same sample composition.

The recommended sample size is 200 consumers for the first tried product with boosts for users of each brand tested (100+) if required. When there are no subgroups to be analysed the Category user sample for the first tried product can be reduced to a minimum of 150. The same respondents are placed with the second product. The order of the products must be rotated to reduce order bias.

[Please indicate deviations from above mentioned sampling]
[Please specify sub groups, if any]

8. Target Group and Recruitment

The selection criteria are strongly linked to the test objectives and target consumers. Typically consumer selection for a quantitative study includes the criteria in the table below.

Criteria per product and region based on the test objectives, test products and target consumers of the product and brand: [Please adjust table for category/sub-category/brand]

	Selection Criteria or Quota’s
- Age	
- Gender	
- LSM (where available)	
- household decision maker and purchaser of product/brand	
- no conflicting interests (e.g. working for a business with conflicting interests to Unilever) → security screening	
- product usage	
- frequency of product usage	
- brand and format usage	
- allergies & diet	
- Other criteria to consider:	
- size of household	
- children in household	

PLEASE NOTE: the test design, sample, questionnaire and recruitment should remain as stable and consistent as possible over time to allow comparison to past results.

<u>Penetration levels</u>

The briefing clearly needs to state what the penetration level is. This should be based on past tests or on household panel data (WorldPanel, Nielsen, etc.)

9. Product

The choice of Unilever products are defined according to Scope. These will approximate products in the UEx top 100 list.
The choice of competitor products is defined by the BPS (brand performance standards), where the competitive set for each brand (platform / sub-brand) is defined. The focus will be on benchmarking our product against the key competitor, not necessarily against the source of growth.
Benchmarking is a test of our product versus the same product / variant of the key competitor (as set in BPS).

Standard Competitive Benchmarking

Standard Competitive Benchmarking must be conducted with product purchased in stores or markets where local consumers would shop for it. This includes both the competitive product and the Unilever product. There should be at least 2 different batch numbers for the competitor and for our product in any given test. This will reduce the impact of anomalies from 'rogue' batches. And for Foods products especially it is important to have similar "Best Before" dates.

The product test for the Unilever product and the benchmarking product need to be conducted in one time period. For example, it is not allowed to test the Unilever product first, and then test the competitor product at a later time.

Benchmarking for Renovated Products

Benchmarking for Renovated Products may use direct-from-factory Unilever product that has been held for an 'aging' period to be determined by the VP R&D for each Category. The Unilever product must come from the real production line, not from a pilot plant. Competitor products, if necessary, should still be purchased from stores or markets where local consumers would shop for it.

Test results and global reporting will need to clearly identify direct-from-factory product to distinguish these tests the in-store product.

Results of these tests will become part of the official Benchmark reporting if they are commissioned for launch in the same or following year, but within the database we will record that the product is from factory trail and when it is in market. *These products will need to be retested with proper in-market samples as soon as the product is established in market (ideally around a year after launch).*

Sourcing Guidelines

2010.05
Benchmarking Re

The product purchase checklist ensures the purchase of the right (amount of) products and batch numbers:

Checklis_product_pu
rchase_v.1.0.doc

Product purchase and de-branding

Unilever is responsible for product purchasing and de-branding of the products. When requested, the Research Agency can support Unilever in purchasing the products, only when a Unilever representative is accompanying the Research Agency to the store and checks that the product purchased is correct. Unilever provides the Research Agency with clear instructions on e.g. batch numbers and shelf life and defines clear responsibilities on who will check and approve the products before the fieldwork starts.

	Instructions
Labelling	[irreducible coding, letters and numbers]
- Product code	
- Batch number	
Text on packs: * - Nutritional labels	
- Ingredients	
- Cautions / Warnings	
- Disclaimers	
Storage	
Preparation	
Quantity respondent has to test	
Items needed to execute the test	[very detailed information on specifications of materials]
Instructions for staff	
[please specify]	

* For Unilever Foods product testing: please provide the Category Nutritionists with this information, including a photo if available for nutritional benchmarking.

Product handling and delivery

It is important to follow all relevant product handling and safety clearance protocols for any products tested with consumers.

If required, a technical protocol, including inspection, clearance, etc. needs to be developed by Unilever (if not already available).

[Please provide specifications in table or separate document:]

The Product handling checklist explains how products should be stored and treated during research:

Checklist_product_handling_v.2.0.doc

10. Respondent Communication

Respondent questions

When applicable, Unilever or the Research Agency need to provide an information telephone number that respondents can call in case of questions, complaints, etc. With blind testing, the Research Agency local office conducting the fieldwork is the primary contact. The Research Agency's telephone number should be placed on the blinded package if possible and should be mentioned in the respondent letter.

In several countries it is mandatory to have the key information, such as product usage, product ingredients, emergency number, etc. on pack (for further details and what goes onto the pack, see Appendix 5). For products which because of their nature and the way they are used (for example, hair care products stored in wet atmospheres) it would be prudent to also have an accompanying sheet of information in case the original label becomes unreadable.

Important Note:
The Unilever Care Line has no knowledge of which products we are testing, and whether the consumer is trying out a Unilever or Competitor product. At placement, consumers should be instructed that if they experience any adverse reaction to the product to seek medical advice immediately. Where required by law, respondents should be given a telephone number to contact if they experience any adverse reaction to the product under test. This is a rare but extremely serious event, and appropriate procedures should be established within the company to deal with it.

11. Questionnaire

The approved global questionnaire template for the category or sub-category is the starting point for all benchmarking studies. The global questionnaire contains core product attributes for each category or sub category. Please ensure that the relevant BPS attributes are included in the questionnaire.

Purchase intent is not allowed to be part of the questionnaire. It is the decision of the PECLT that purchase intent is meaningless (and in fact distracting) at this stage.

The master questionnaire must be in accordance with all guidelines and product testing protocols and will be reviewed and approved by the Research Agency category Lead and the Research Agency Program Management. Local offices/regional offices can add /adjust the questionnaire to local markets (maximum 10% of questionnaire). Additions should be placed at the back-end of the questionnaire, to ensure a correct order. If questions interfere with the standard questionnaire order, an approval from the Research Agency Program Management is required.
Note that the number of attributes included during questioning should be maximum 20-25, with at least 10-15 consistent by category.

The batch number of the products tested must be in the questionnaire to be able to deduce possible differences in results.

a) Standard outline of master benchmarking questionnaire (Recall):

	Question	Scale	Scales	Reverse scale for North America
1.	Confirm usage of test product			
2.	Overall opinion	7-point	1.Very poor 2.Poor 3.Neither poor nor fair 4.Fair 5.Good 6.Very good 7.Excellent	7.Excellent 6. Very good 5.Good 4.Fair 3.Neither poor nor fair 2.Poor 1.Very poor
3.	Likes	Open ended		
4.	Dislikes	Open ended		
	Attribute ratings (aligned with BPS)			
	Key overall uni-polar attributes	7-point	1.Very poor 2.Poor 3.Neither poor nor fair 4.Fair 5.Good 6.Very good 7.Excellent	Reverse scale
	Other uni-polar attributes	5-point	1.Disagree strongly 2.Disagree a little 3.Neither agree nor disagree 4.Agree a little 5.Agree strongly	Reverse scale
	Bi-polar attributes	5-point	*Just-right scales* 1.Much too [strong] 2. A little too […] 3.Just right 4.A little too […] 5.Much too [weak]	Reverse scale
	Optional: Specific diagnostics	5-point	Just-right scales, see 'bi-polar attributes'	Reverse scale

The recall questionnaire should be identical for the first and second product usage.

b) Preference question

After the recall interview for the second product has taken place, we ask the consumer a preference question. This is only on overall opinion of the product. This question will have three different response options:

Taking everything into consideration, which of the two products did you prefer overall?

- I preferred the product I tried first
- I preferred the product I tried second
- I have no preference

c) Brand recognition question

The following brand recognition questions must always be included in the questionnaire at the end of the survey:

1. Did you recognise the product we gave you to try?
   No, I did <u>not</u> recognise the product GO TO NEXT SECTION
   Yes, I did recognise the product GO TO NEXT QUESTION

2. What did you think it was?
   DO NOT READ NOR SHOW LIST
   Pre-coded list as appropriate to study

<u>d) Code list</u>

* Like and dislike questions are part of the master questionnaire, but coding of these questions is only applicable when:

- the product is perceived positively: coding of the likes to be able to define communication routes
- the product is perceived negatively: coding of the dislikes to be able optimise the product

In the proposal, coding per open-ended question is <u>an optional cost item</u> in the investment overview. The cost for this should be itemised in the proposal and the Unilever team must decide whether or not to code before the test begins. If the Unilever project owner decides to change this later, the Research Agency must receive approval for any additional costs before beginning the extra work.

The Research Agency is responsible for coding open-ended comments. If already available, Unilever can provide a master category code list, including all codes for the category, including sub-categories/brands. The code list is translated into all required languages.

<u>e) Translations</u>

Translations are provided by Local Research Agency Teams and are approved by Local Unilever. During translation the <u>contents</u> of the questionnaire <u>cannot be changed</u>.

12. Timing

The length of the product test period should be set to ensure the respondent has sufficient time to assess the product adequately. The specific test length will depend on frequency of usage of the product (and this will differ by category).

For products that are used over time (e.g. fabric conditioner) the test length will normally be 1-3 weeks per product tested. The allocated time should take into consideration the number of desired uses necessary to see the benefit.

Each category should set its own rules on a country to country basis as appropriate for the product concerned.

For products that are consumed fully on a single occasion (e.g. some food products), a single pack should be provided and the recall interview will normally be fixed for about one week later. If tested in-home, the test length should normally be set in units of a week to ensure adequate representation of each day of the week. For product tested in a Central Location test [please add category's protocol].

For products with a 'carry over' effect, that occurs over several weeks (as in the case of a colour protection benefit provided by a detergent or therapeutic effects of a shampoo), the length of the test may need to be extended and a sequential-monadic design might not be suitable.

[Please indicate: test length per sub category (or brand) per country]

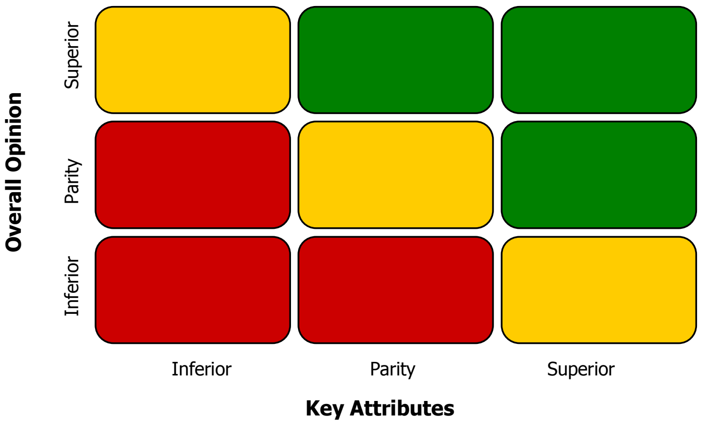
Sub category (or brand)	Country	Test length
		# weeks*

* Including each day of the week.

13. Brand Performance Standards / Action Standards

The UEx benchmarking cross-category KPI for 2012 is 50% win, 40% parity and 10% loss. By end of 2011, we aim to get at least 50 wins cross-category. Each category will have their own category target to come to the 50 wins.

Wins are determined according to the 3x3 structure below. This is set on Overall Opinion x Key attributes. Only those test results that are placed in the green boxes constitute a win. For those tests where key attributes were not recorded (period before 2011), the key attribute default value is set on "parity".



Key attributes (technical + blind) that drive Claimable Functional Superiority

The rule for assessing the Key Attribute value is as follows:

- Only if all key attributes are at win (95% sign.), the key attribute score is “**superior**”
- If more than half the key attributes are at loss, the Key Attribute score is “**inferior**”
- All other cases are “**parity**”.

Action standards must be set for the Unilever product to be superior to competition on Overall Opinion and/or key attributes at the 95% level of confidence (two-tailed test) (applied to mean score).

Action Standards must be agreed by all parties prior to the test and must be written down as part of the Market Research proposal. Action Standards will include an overall measure (such as ‘Overall Opinion’ mean score) as well as some key attributes (2-5 attributes as per category agreed list). The key attributes must be part of the BPS and be assessed in every test. Key attributes are a sub-set of BPS. They are technical + blind consumer test attributes set on win (see also background).

Note that in some instances, a key attribute can consist of two underlying key attributes, reflecting the same consumer benefit. In these instances, at least one of the underlying key attributes has to be a win, whereas the other attribute(s) can not be a loss.

14. Significance Testing

For Competitor Benchmarking, Unilever mandates the 95% confidence level (two-tailed test) t-test for key action standards. The reason for this is that it establishes clear likelihood that consumers will notice the difference in the marketplace.

The combined 1st and 2nd position data (aggregated sequential results) will be reported for studies where there are no order effects. If there are no order effects, that should be noted on the Action Standard Chart.

Order effects are defined as situations where people may judge the second product relative to the first one rather than an internal norm.

To ensure that any potential order effect - if at all present - is not big enough to cause a change in conclusion, the following steps are required:

4. Compute the difference between the Unilever product and a competitive product in the First position. Call that difference (A).
5. Then compute the difference between the Unilever product and a competitive product in Total. Call that difference (B).
6. If A and B have different signs (one is positive and one is negative) then one should use the test results on First position only (reporting monadic data). Otherwise, we conclude report Total results (aggregate sequential-monadic data).

Note that this is an effect that needs to be measured on both Overall Opinion AND Key Attributes and needs to be measured for each test. Indicate on the Action Standard chart if an order effect has been found and whether combined or first position results are in the Action Standard using one of the following notes:

3) Note: No order effect was found so the combined results are reported
4) Note: An order effect was found so the monadic results are reported

Using the aggregated data set rather than only the monadic read would strengthen the reliability of the results (as the number of consumers testing a product is doubled, reducing the likelihood of a Type 2 error).

15. Analysis of Results

The required analysis of results is:

- Overall Opinion and Key Attributes on first product tested
- Analysis on the full set for diagnostics
- Results on the preference question

Drivers Analysis on a single-study basis must not be done. Results from such analysis give no further insights and can set teams off in the wrong direction. Links between technical and blind consumer test results should be done on a single-study basis. The purpose of this quality benchmarking is to provide a quick, easy and clear overview with monadic reading on the first product tested, completed with sequential monadic and forced-choice question to have more diagnostics. It is not meant for problem analysis. Separate research/studies will be necessary for that.

Drivers Analysis, correlating attributes from blind tests to overall opinion, is a very useful tool over multiple studies over several competitors and can also be used to guide communication (claims). Correlation between technical (sensory / analytical) data and blind test data (overall opinion or key attributes) should be done on a single study basis to guide product quality improvement. To truly understand the derived drivers, teams will need a larger scale study, such as preference mapping.

To help guide Unilever in the future, please also ensure that you do get the complete data file with results per participant to guide further studies, done by the capability team. This would be an Excel spreadsheet with all participants' scores.

16. Reporting

In order to ensure consistency across categories, a top-line strategic summary of the data using the traffic light reporting format is mandatory, with significance levels set at the 95% level of confidence.

In addition, a one page management summary using the Power Page template must also be prepared and sent as indicated in the Power Page templates..

Power Page Template (CTI/CMI):

REPORT_POWER PAGE_CTI Template_

REPORT_POWER PAGE_CMI Template_

More detailed reporting and recommendations for each individual test should comprise the following:

Introduction

- Field period (exact dates)
- Length of placement
- Products included
- Nr. of interviews per product
- All recruitment criteria and quotas
- Length of questionnaire
- Method
- Recruitment method (door to door (random route), self recruited online data base, pre recruitment …)
- Agency
- Test area/cities
- Sample profile (user/non user, age, water hardness, cities, powder/liquid user, any other quota criteria, extra page with brand users nowadays)
- Weather conditions

APPENDIX - Definitions of Commonly Used Statistics

T – Test
A statistical test used to assess the significance of the difference between two statistical summary measures, typically means.

The test has two components: 1) the numerator is the difference between the two means to be tested for significance and 2) the denominator is a measure of variability, called the standard error of the difference between means, which reflects an expected amount of difference between the means if they differed by chance or due to randomness only. Roughly, a t-test value of 2 or greater, indicating the difference between means is twice the size of the expected variability, is considered significant. Technically, the t-test is a general test of the difference between two sets or distributions of data. The test is sensitive to differences not only in means but also the variation of responses around the
mean as well as the shape of the distributions.

Z- Test
A statistical test used to assess the significance of the difference between two percentages (or proportions).

The test has two components: 1) the numerator is the difference between the two percentages to be tested for significance and 2) the denominator is a measure of variability, called the standard error of the difference between percentages, which reflects an expected amount of difference between the percentages if they differed by chance or due to randomness only. Roughly, a z-test value of 2 or greater, indicating the difference between percentages is twice the size of the expected variability, is considered significant. Technically, the z-test
is a general test of the difference between two sets or distributions of data. The test is sensitive to differences not only in percentages but also the variation of responses around the percentages as well as the shape of the distributions.

Confidence
With reference to statistical testing, the amount of statistical assurance or reliance provided by a statistical test in the assessment of a fact derived from data.

Tails
With specific reference to statistical significance testing, the statistical adjustment made to significance tests to accommodate the nature and direction of differences tested.

If direction of the difference is not important then the researcher only wishes to consider whether there is a difference or not. The statistical test accommodates the possibility that either object may be greater or less than the other.
On the other hand, the nature of difference, that one specific object was better (or worse) than the other, may be very important. The research goal is to assess the chances that a specific object is considered as best (or worst) and as such the direction of the difference is important. In statistical parlance, this is consistent with the use of a one-tailed test