Food Quality and Preference 47 (2016) 3–9

Contents lists available at ScienceDirect

Food Quality and Preference

journal homepage: www.elsevier.com/locate/foodqual

Assessing sensory panel performance using generalizability theory

Paul Talsma *

Givaudan UK Ltd., Kennington Road, Ashford, Kent TN24 0LT, United Kingdom

ARTICLE INFO

Article history:
Received 4 September 2014
Received in revised form 21 January 2015
Accepted 27 February 2015
Available online 12 March 2015

Keywords:
Generalizability theory
Panel performance
Coefficient Φ
Panel reliability
Variance component
Statistical power

ABSTRACT

Generalizability theory provides a framework for assessing panel reliability, both for the panel as a whole and for individual panellists. The variability of the sensory panel scores is split up into products, panellists, replications, and interactions between these terms. Reliability is defined as product variance over total variance. Coefficient *G* only includes terms including product in the denominator and focusses on the ordering of the product scores.

Coefficient Φ is introduced, which includes all variance components in the denominator and focuses on the ordering as well as on the absolute values. It is shown that this latter feature provides important additional information about panel performance.

An algorithm is described which excludes panellists one by one and then evaluates the contributions of each panellist on total test reliability. The focus is on changes in the coefficients and variance components after removal compared to before, allowing for an in-depth evaluation of the performance of individual panellists. When coefficient Φ increases with an amount deemed relevant (0.05 on average for all attributes, or 0.10 for a single attribute) after removal of a panellist, the panellist qualifies for exclusion from the statistical analysis. The total number of panellists to be excluded is limited to a maximum of 20% of the panel size.

It is shown that a statistical power calculation is a useful addition to a reliability analysis by checking if panel discrimination meets a pre-set standard.

It is explained how the reliability algorithm and the power calculation can be implemented using MS Excel.

The common criteria for assessing panel performance: discrimination, consensus, and repeatability, are defined in terms of generalizability theory variance components. Discrimination focuses on maximising product variance, consensus on minimising all components containing panellist, and repeatability on minimising all components containing replicate.

It is discussed how reliability results obtained using this methodology can be used for panel management.

Two examples from studies carried out by the Givaudan sensory panel in Ashford, UK, are given.

1. Introduction

This article builds upon a recent publication in Food Quality and Preference (Verhoef, Huijberts, & Vaessen, 2015) where generalizability theory (Brennan, 2001) is proposed to evaluate panel performance. They give a good introduction into the background and subject matter. At Givaudan we have been using generalizability theory to assess panel performance for 4 years. This article describes three additions to the framework suggested by Verhoef et al., and gives two examples of applications in research studies.

In generalizability theory (*G* theory), the sensory panel data variance is split into its components: products, panellists, replications, and interactions between these. *G* theory can be used for reliability analysis (a *G*-study) or for decision making (a D-study). In a *G*-study on the one hand, the focus is on obtaining estimates of variance components and using these to assess reliability of the raw data scores. In a D-study on the other, the data are collected for the specific purpose of making a decision. At Givaudan, we want to make decisions about products or fragrance samples, like for example which one has a higher fragrance intensity than the others. So for us the D-study framework applies. We use the product means for decision making. We therefore focus on the reliability of the product means rather than of the raw data. The variance components of these are derived as indicated in Table 1.

* Tel.: +44 (0) 1233 644090.
E-mail address: Paul.talsma@givaudan.com

http://dx.doi.org/10.1016/j.foodqual.2015.02.019
0950-3293/

Table 1
Variance components in a sensory panel study: derivation and reliability coefficients.

Variance component	Derivation	Included in denominator	
		G	Φ
Product	$\sigma^2_{\text{product}}$	√	√
Panellist	$\sigma^2_{\text{panellist}}$/number of panellists		√
Replicate	$\sigma^2_{\text{replicate}}$/number of replicates		√
Product by panellist	$\sigma^2_{\text{product * panellist}}$/number of panellists	√	√
Product by replicate	$\sigma^2_{\text{product * replicate}}$/number of replicates	√	√
Panellist by replicate	$\sigma^2_{\text{panellist * replicate}}$/(number of panellists * number of replicates)		√
Product by panellist by replicate	$\sigma^2_{\text{product * panellist * replicate}}$/(number of panellists * number of replicates)	√	√

Table 2
Algorithm for evaluating the panel performance of individual panellists.

Step 1	Reliability whole panel (full set) G and Φ	All panellists
Step 2	Remove one panellist and recalculate	For all panellists
Step 3	Does removing the panellist increase Φ with a relevant amount? If yes, remove	For each panellist
Step 4	Recalculate coefficients for reduced set	Reliable panellists

2. Calculation

We estimate the respective variances in Table 1 using SAS PROC VARCOMP using minimum norm quadratic unbiased (MINQUE) estimation (Hartley, Rao, & LaMotte, 1978). The MIVQUE0 method in SAS (METHOD = MIVQUE0) produces unbiased estimates that are locally best quadratic unbiased estimates given that the true ratio of each component to the residual error component is zero. Negative estimates, which sometimes occur and are caused by variability in the data or outliers, are set to zero.

We normally present the variance components as percentages adding up to 100. This gives good insight into possible causes for lack of panel reliability.

Reliability is defined as product variance/total variance, where the total variance is calculated as the sum of the variance components indicated in Table 1 for reliability coefficients G and Φ. Coefficient G is the usual reliability (generalizability) coefficient and is called univariate quality index by Verhoef et al. (2014).

2.1. Coefficient Φ

Coefficient Φ has been proposed by Brennan (2001) and we consider it to be a useful addition. Unlike G it takes all possible sources for lack of reliability into account. Whereas G focusses on the ordering of the product scores, Φ focusses on both the ordering and the absolute scores. Φ is always $\leqslant G$.

If, for example, we would find that for our panel $G = 0.91$ and $\Phi = 0.50$, we can see that by and large the panel have excellent agreement about the ordering of the product scores across replicates, but the low level of Φ shows that the following three variance components are relatively large in combination: the main effect of panellist, the main effect of replicate, and the panellist by replicate interaction. With some straightforward arithmetic it can actually be derived that the combined variance of these three components equals 90% of the product variance ($100\% * [1/\Phi - 1/G]$). For a more informed view, the actual variance components should be studied. Their interpretation will be explained in detail below.

2.2. Algorithm

Coefficients G and Φ are useful to evaluate performance of the panel as a whole, but do not give insight in the performance of individual panellists. For this we have used an algorithm of which the steps are explained in Table 2.

We start with calculating coefficients G and Φ of the panel as a whole, and then proceed with removing one panellist at the time and recalculating the coefficients. Coefficient Φ with the panellist removed is compared to Φ of the whole panel, and if the reliability increases with a relevant amount when the panellist is not included, then the panellist is considered to have a negative effect on the overall test reliability. We use as criteria for relevance: either Φ increases with 0.05 or more on average for all assessments together, or Φ increases with 0.10 for one single assessment. For the second criterion it is also checked if Φ does not decrease on average for all assessments together.

Panellists thus identified have a negative effect on test reliability which may be due to several reasons like underperforming on the day or inexperience/lack of training. This line of reasoning only applies, however, if the occasional panellist is not performing on the day and the majority of the panel is. For this reason, we only use the reduced set for reporting when no more than 20% of panellists have been removed from the analysis. Otherwise, the full set is used.

The cut-off level at 20% has been agreed upon within Givaudan. This is the criterion that sensory test leaders are using and have been using for a long time, because it fits with their perception of an appropriate boundary between the occasional panellist not performing and the majority of the panel not performing. It is also being used in the Flavours division of Givaudan, and has been defined independently by them.

The test leader should always check the data and evaluate if they agree with the decisions taken by the algorithm. Having the algorithm in place does have the advantage, however, that the criteria for panellist removal and which (sub)set to use for the report are all specified in advance before the test is done.

2.3. Discrimination: Statistical power analysis

Reliability coefficients can be very low (close to zero) while the panel has performed adequately. The reason for this is the definition of reliability as product variance over total variance. When the products are very similar, product variance may be very low, leading to low values for the coefficients.

For this reason, we have introduced the approach of supplementing the reliability analysis with statistical power analysis.