Food Quality and Preference 49 (2016) 1–10

Contents lists available at ScienceDirect

Food Quality and Preference

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

On the analysis of Rate-All-That-Apply (RATA) data

Michael Meyners [a,*], Sara R. Jaeger [b], Gastón Ares [c]

[a] Procter & Gamble Service GmbH, 65824 Schwalbach am Taunus, Germany
[b] The New Zealand Institute for Plant & Food Research Limited, 120 Mt Albert Road, Private Bag 92169, Victoria Street West, Auckland, New Zealand
[c] Facultad de Química, Universidad de la República, Gral. Flores 2124, C.P. 11800 Montevideo, Uruguay

ARTICLE INFO

Article history:
Received 8 September 2015
Received in revised form 6 November 2015
Accepted 9 November 2015
Available online 11 November 2015

Keywords:
Rate-All-That-Apply (RATA)
Check-All-That-Apply (CATA)
Randomization test
ANOVA
PCA
CA

ABSTRACT

Rate-All-That-Apply (RATA) is a variation of the more widely used CATA question format. For a pre-specified list of terms, consumers indicate whether they apply to a given product, and if they do so, to rate their intensity. For example, a 3-pt scale may be used with anchors 'low', 'medium' and 'high' or a 5-pt scale anchored at 1 = 'slightly applicable' and 5 = 'very applicable.' Given the hierarchical nature of the task and the non-normal distribution of the intensity data, it is not obvious how to analyze RATA data appropriately. In the present work we suggest interpreting RATA data as 4- or 6-point scales, considering a missing check for any attribute as a score of 0. Based on that, randomization tests were applied to investigate potential product differences. We show that the null distribution of these tests for RATA data coincides in practice with the one from classical parametric tests derived from an ANOVA context. Consequently, using the common *F*- and *t*-tests provides a valid and easy analysis of RATA data. In four consumer studies, tests for product difference usually gave the same results when RATA data were analyzed with ANOVA and Cochran's *Q* test. Graphical display of the data was found to be very similar based on Correspondence Analysis (CA) (treating RATA as CATA data) and Principal Component Analysis (PCA) (treating RATA as continuous data).

1. Introduction

1.1. Background and research aim

The application of CATA questions (Check-All-That-Apply) is gaining popularity as a quick and easy approach to obtaining consumer-based sensory product characterization (Ares & Jaeger, 2015). These questions consist of a list of terms from which consumers have to select all those that they consider applicable to describe the focal sample. CATA questions provide valid and reliable data when applied to a wide range of consumer products (Ares et al., 2015; Jaeger et al., 2013; Meyners & Castura, 2014).

One of the limitations of CATA questions is that they do not allow a direct measurement of the intensity of the perceived sensory attributes, which could potentially hinder discrimination among products that have similar sensory characteristics, but slightly differ in the intensity of those characteristics. For example, *sweet* may apply to all the focal products in a test. CATA questions capture this information, but it is less obvious that they are able to capture differences in sweetness intensity between the focal products. Approaches for jointly obtaining CATA and intensity-based responses are emerging as a consequence, as exemplified by Reinbach, Giacalone, Ribeiro, Bredie, and Frøst (2014).

Similarly motivated, Ares, Bruzzone, et al. (2014) introduced RATA (Rate-All-That-Apply) questions. In this variant of the CATA question format, when a consumer assesses a sample, s/he first decides if a given product attribute applies or not. Only if it applies, s/he will also rate its intensity/applicability. Two rating scale variants were presented: a 3-pt intensity scale with anchors 1 = 'low', 2 = 'medium' and 3 = 'high' and a 5-pt applicability scale with end-point anchors 1 = 'slightly applicable' and 5 = 'very applicable.' Compared to CATA questions, the use of a RATA variant was found to increase the number of attribute terms selected for describing samples and led to a small increase in the percentage of terms for which significant differences among samples were identified.

In view of the recent development of RATA questions, little attention has been directed to the analysis of such data and it remains unknown how decisions regarding analysis influence results of sensory product characterizations from this question format. The current research contributes to closing this gap by providing recommendations for analysis of RATA data. Four consumer studies with different product categories (3–7 samples per study) were conducted in New Zealand and Uruguay (50–56 participants per study) (see Table 1 for summary details). For the obtained RATA data, we

* Corresponding author.
E-mail address: meyners.m@pg.com (M. Meyners).

http://dx.doi.org/10.1016/j.foodqual.2015.11.003
0950-3293/

Table 1
Number of significant overall comparisons among samples obtained at level 5% via ANOVA and number of significant pairwise comparisons at level 5% obtained with subsequent *t*-tests and via treating RATA as CATA data (Cochran's *Q* test). The last column provides the *p* value for the binomial test investigating whether ANOVA or RATA as CATA yielded a higher number of significant pairwise comparisons.

Study	Number of significant comparisons									
	Number of products/ RATA terms	Overall comparisons				Pairwise comparisons				
		Both	None	ANOVA only	RATA as CATA only	Both	None	ANOVA only	RATA as CATA only	*p* Value
1. Fruit cake*	5/12	11	0	1	0	57	39	20	4	0.002
2. Apples*	4/16	10	5	1	0	32	56	7	1	0.070
3. Peanuts*	3/12	9	1	2	0	17	14	4	1	0.375
4. Milk desserts#	7/18	5	8	3	2	15	275	61	27	<0.001

Notes. *3-pt RATA scale with anchors 'low', 'medium' and 'high'. #5-pt RATA scale with end-point anchors 1 = 'slightly applicable' and 5 = 'very applicable.' The number of significant pairwise comparisons is equal to the number of RATA terms (*N*) when considering overall comparisons. For pairwise comparisons, the number of comparisons is equal to $N \times P \times (P-1)/2$, where *P* is the number of products in a study.

investigate the appropriateness of using *F*- and *t*-tests based on ANOVA (Research Question 1). If this is shown, we compare results from this parametric analysis to when RATA data are treated as CATA data (i.e., ignoring intensity/applicability ratings, which are all treated as taking the value 1) and investigate their relative performance (Research Question 2). Before presenting methodological details pertaining to data collection, the proposed approach to the analysis of RATA data is further explained and justified.

1.2. Development of a randomization test to investigate whether F- and t-tests based on ANOVA are appropriate for the analysis of RATA data

By design, a RATA task has two parts. First, respondents decide if an attribute applies or not. Only if they decide it applies, they rate on a scale how intense/applicable the attribute is. Despite the stepwise setup, it seems worthwhile ignoring the hierarchical nature of RATA data and treating the scores as they are, while introducing the value 0 in case the attribute was not considered applicable to the focal product. By doing so, a 3-point RATA scale turns into a 4-point scale (0 = attribute not selected; 1, 2, 3 = attribute selected and intensity rated as 'low', 'medium' and 'high', respectively). Similarly for scales with a different number of points, whereby, for example, a 5-pt RATA scale turns into a 6-pt measurement scale. This way of handling the data leads to the question of whether RATA responses can be analyzed by classical parametric methods, i.e. analysis of variance (ANOVA) and *t*-tests. For sensory profiling data on a limited scale and in particular with an inflated number of zeros (i.e. with deviations from normality), Kunert, Meyners, and Erdbrügge (2002) show that the *F*- and *t*-test typically employed in ANOVA are usually quite robust against this deviation from normality. Against this background, we investigate whether the parametric tests associated with an ANOVA might be used for the analysis of RATA data. As ANOVA is widely available and the most commonly used tool to investigate sample differences, this would offer an easy approach to appropriately analyze RATA data (*F*-tests are used for comparisons of 3 or more products, while *t*-tests are used for pairwise comparisons).

It is worth noting that the approach assumes the 0-appended RATA data to have equal distances between adjacent categories. This is a limitation of any analysis based on mean values and alike, yet an assumption that is often made in similar situations and that has been shown to provide useful results. Whether this assumption is reasonable remains an open research question, though, and cannot be answered with the data at hand.

As we will show in the sequence, RATA data does usually not meet the assumptions formally required by an ANOVA. Yet, to investigate if *F*- and *t*-tests based on ANOVA are nevertheless reliable analyses of RATA data (RQ1), a gold standard is needed against which to compare the performance of the parametric tests. The natural candidate is an appropriate randomization test (cf. Edgington & Onghena, 2007; see Meyners & Pineau, 2010, and Meyners, Castura, & Carr, 2013, for a description in the context of sensory data). The benefit of this class of tests is that its validity does not depend on parametric assumptions, but is warranted by the experimental design, most notably by the (possibly balanced or stratified) randomization of products to evaluations, as is standard practice in most sensory/consumer studies.

To execute the comparison, we propose the following: If, in a given study, the focal products are not different (i.e. under the null hypothesis), the evaluations of the samples are not dependent on the product, as they are (at least perceptually) equivalent. In this instance, we can randomly permute the allocation of products to evaluations within the same constraints as for the original allocation. This should not systematically increase or decrease a reasonable test statistic, if the null hypothesis is true. If the null hypothesis is not true and the products are different, a systematic trend towards smaller (or larger) values will be observed. This creates a decision rule to judge on statistical significance of differences between products. To compare this approach versus the corresponding parametric test, we will take the corresponding test statistics and record it for a substantial number of re-randomizations of the product-to-evaluation allocation. This gives the null distribution for the randomization test and the respective test statistics. We can then compare this null distribution from the randomizations with the corresponding parametric null distribution. If these distributions are very similar, the tests will give the same (or at least very similar) results. Note that if the *p* value is around 5%, it should be quite precise to assure the correct test decision is made at a significance level of 5%. However, if the *p* value is determined to be 30%, it does not really matter much for interpretation (and not at all for the formal statistical test) whether it is truly 30%, or rather 27% or 33%, say. As significance levels higher than 10% are rarely used in sensory and consumer research, it usually suffices if the distributions are very similar in the extremes, i.e. below 10% or above 90% of cumulative probability. If the distributions coincide in that range, the test decisions (i.e. to reject or retain the null hypothesis) derived from the two distributions are identical, even if the distributions differ in other parts.

It is worth mentioning that the approach taken is very similar to those used by Eden and Yates (1933), Fisher (1935), and Pitman (1937a, 1937b, 1938; see also David, 2008, for a brief overview). These pioneers have used the randomization approach to validate parametric tests like the *F*- and *t*-tests in simple and more complex models. Their work provided empirical evidence for the validity of the parametric tests, which then allowed efficient performance of statistical testing, in particular prior to the broad availability of personal computers.