PAIRED COMPARISONS OF STATEMENTS: EXPLORING THE IMPACT OF DESIGN ELEMENTS ON RESULTS

Roberto Furlan[1], Graham Turner

Kantar Health, Epsom, UK

Received: 10th September 2013/ Accepted: 27th January 2014

Abstract. *Paired Comparisons of Statements (PCS) analysis, a technique that has its roots in the law of comparative judgment introduced by Thurstone (1927), has gained a significant increase in popularity in market research in recent years. PCS represents a valid and better alternative when collecting preference measurements, being scale-free and providing more differentiation when measuring attribute importance than standard rating scales. With growing popularity there is a clear need to better understand the potentialities and limitations of PCS. While some preliminary work has already been done (Corradetti and Furlan, 2006), there are still several unexplored areas, in particular regarding the impact of the key elements in a PCS design on the accuracy of results. With this paper we try to understand to what extent the number of versions, the number of tasks, the model/ scale, and the number of respondents impact on the results.*

Keywords*: Paired Comparisons of Statements, PCS, comparison, hierarchical Bayes.*

1. INTRODUCTION

Companies constantly seek to enhance customer satisfaction and retention by improving the overall quality of a product or service. To do so, managers must focus on enhancing particular attributes of the product or service, those with the potential greatest impact on customer satisfaction. However, identifying such key characteristics can be challenging and a key step is determining the value customers attach to the different features. The market researcher has several tools in the arsenal to assess such value. Among these, the most popular metrics are traditional approaches such as ratings, rankings, and constant sum. However, in the last decade trade-off approaches such as Maximum Difference Scaling (MDS) (Louviere and Woodworth, 1990; Finn and Louviere, 1992) and Paired Comparisons of Statements (PCS) (David, 1988; Corradetti and Furlan, 2006) have become rather popular among

[1] Roberto Furlan, email: roberto.furlan@kantarhealth.com; roberto.furlan@gmail.com

market researchers due to their advantages over the more traditional techniques. In the literature, one can find plenty of theoretical and empirical studies dedicated to these individual approaches and also some works involving a comparison of different methodologies. For instance, Chrzan and Golovashkina (2006) conducted a study to test six different types of importance metrics including traditional approaches (i.e., ratings and constant sum), MDS, and three other less popular methodologies; Jaeger *et al.* (2008) compared MDS to preference ratings; Madansky (2010) considered PCS, MDS, and preference rankings. All these studies were based on empirical results.

Paired Comparisons of Statements is a discrete choice model that has its roots in the law of comparative judgment presented by Thurstone (1927) and that has been extensively described by David (1988) and more recently by Corradetti and Furlan (2006). To date, it is widely used to collect and scale preference measurements through a structured research questionnaire. The researcher defines a set of items (usually statements, messages, product features, service characteristics, options in a decision, etc.) and assumes that there is some underlying subjective dimension, such as extent of preference, degree of importance, degree of credibility, extent of appeal, impact on prescription (for medical products), impact on purchasing, etc. In the PCS approach, the ultimate aim is to measure the location or position of the set of items on that dimension. These locations are estimated through an algorithm that provides a set of utilities, with one utility score associated to each item.

In a PCS exercise, survey respondents are repetitively shown subsets of size two of the possible items (each subset is also referred to as a PCS task). In its simplest setup, referred to as *short paired comparison of statements*, the respondent is asked to choose the preferred item (or the most credible, important, appealing, etc.) from each subset. As the resulting data are quite poor from both a psychological and a statistical points of view, the researcher often prefers to ask the respondent to also indicate the intensity of the preference in what is called *graded paired comparison of statements* model. In the graded version, the two items are usually presented horizontally and a scale is presented underneath. In both the short and graded version the researcher might decide to include a neutral point for indicating 'no preference', useful when one does not want to force respondents to make a choice towards one of the two items.

To some extent PCS is a valid and popular alternative to self-explicated models. In this class of models, respondents would directly rate or rank the elements or allocate a number of points among them. With a rating approach, survey respondents are presented the features individually and asked for their evaluations. While this exercise is straightforward and requires little time and effort, it does not

explicitly capture priorities and results might suffer from lack of differentiation (e.g., everything emerges as being important); in addition, the scale suffers from scalar inequivalence issues (i.e., due to response style and cultural and personal background differences there might be differences across respondents in the usage of the scale - Louviere and Flynn, 2011; Sawtooth Software, 2007). All these drawbacks might compromise the correct interpretation of the results and thus the actionability.

The ranking approach would not present these issues, however rank evaluations imply an ordinal scale, while some researchers prefer to work with interval or ratio scales because of their statistical properties. Similarly to the ranking approach, the popular constant-sum allocation, an approach requiring respondents to divide a limited amount of resources across a number of elements, captures priorities quite well and the scale is not affected by the inequivalence issue. However, with a large number of (six or more) elements, it becomes very difficult for the respondent to effectively allocate scores among all of them, thus limiting the applicability of this approach to only the smallest sets (Srinivasan and Wyner, 2009).

In this context, PCS represents a valid approach to collect preference data, as it is based on a trade-off approach rather than a self-explicated one. It is a rather simple exercise, usually requiring an acceptable effort from respondents, it is simple to execute, it can handle many elements, it provides results that are empirically consistent with more complex ordering tasks, and produces reasonable differentiation in the results which appear to be on a convenient ratio scale. Probably, the most important property is that it measures all the items on a common scale, thus addressing the scalar inequivalence problem characterising the way respondents use rating scales, arising mostly from differences in response styles and cultural differences (Cohen and Neira, 2003; Steenkamp and Hofstede, 2002).

PCS is known with different names: Method of Paired Comparisons (David, 1988), Multiple Paired Comparisons, Trade-Off of Statements (Corradetti and Furlan, 2006), Scalar Conjoint (Department of Trade and Industry, 2002; GfK, 2010). The last name reminds that among some users, PCS is erroneously referred to as a conjoint analysis technique. Conjoint analysis is in fact defined as “any decompositional method that estimates the structure of a consumer’s preferences given his/her overall evaluations of a set of alternatives that are pre-specified in terms of levels of different attributes” (Green and Srinivasan, 1978). While both PCS and conjoint models require an experimental design and both aim to estimate the structure of a respondent’s preferences, PCS is not based on a set of alternatives characterised in terms of levels of different attributes like conjoint analysis. In fact, PCS works with dichotomous items that can either be included or absent in any PCS

tasks, and not with multi-attribute profiles (characterised by a number of attributes, each described in terms of one level). Because of its model specification, PCS is appropriate for research problems focusing on the contrast between items in terms of preference or importance, rather than being based on the additive effect across multiple items like in a conjoint setup. Nevertheless, the label 'conjoint' associated to the PCS exercise remains popular in market research, as from a marketing point of view it is particularly convenient to exploit clients familiarity with this term.

While PCS is considerably different from a conjoint model, there are lots of similarities with the MDS model. In this model, the researcher defines a set of items (as in the PCS model) and survey respondents are repetitively shown subset of the possible items and are asked to indicate the best and worst items (or most and least credible, important, appealing, etc) from each subset. Due to the similarities, MDS is sometimes considered an extension of the well-established short PCS (Sawtooth Software, 2007). For a better understanding of the MDS model and its main characteristics the reader may consider the work done by Louviere and Woodworth (1990).A key aspect to be considered when preparing a PCS exercise is designing an appropriate series of choice sets that would allow estimating the items' preferences. All sets across the sample respondents characterise the PCS experimental design, which can be either complete or incomplete. The *complete* case requires C_2^P pairs, where P is the number of items considered for the study. It is easy to deduce that in the complete PCS setup, the number of pairs required is extremely large, even with a relatively small number of items P. Consequently, no large application can be realised using the complete PCS method. By using an *incomplete* procedure and by determining the number of times that each item will appear in the design, it is possible to substantially reduce the number of pairs required. This approach uses some important results from design of experiments. In particular, it takes advantage of some incomplete block designs. There are two broad classes of incomplete block designs to reduce the number of pairs required: cyclic designs and group-divisible designs. Jackson and Shenker (1981) suggested employing a cyclic design to reduce the number of pairs required and to determine the pairs to be presented. Cyclic designs (CD) are incomplete block designs that do not satisfy balance criteria, but are quite easy to construct. They tend to produce estimates of items differences with similar precision and can be highly efficient (John *et al.*, 1972). In a CD for a PCS exercise with P statements, there are as many statements as blocks (the size of each block is 2) and each statement appears twice in the design. When P is an odd number, there are $C(P, 2) / P = (P-1)/2$ possible CDs. If P is even, then there are $(P/2) - 1$ CDs, plus another one which presents a symmetric structure and therefore may count as half CD. These designs are never balanced because each

statement appears only once with two other treatments, and never with any other treatment. Group-divisible designs are an important class of partially balanced incomplete block designs with two associate classes. Dean and Voss (1999) describe in detail the properties of these designs and also how to build them.

The simplest way to analyse PCS data is through a logit model (Corradetti and Furlan, 2006). Let P be the set of items in the experimental design and T the set of PCS tasks to be evaluated. Each task $t \in T$ is assessed through a preference score assigned to one of the two presented items. The evaluation for task t is stored onto $\mathbf{y}_t$, an interval-scaled variable that can assume values in the range [–s, +s] where s is a positive integer set by the researcher. A negative value for $\mathbf{y}_t$ indicates a preference for the first/left item in the task while a positive one indicates a preference for the second/right item. The indifference for either of the two items is expressed by $\mathbf{y}_t = 0$ and it can be made available or not to respondents. While in the short paired comparison setup $s = 1$, in the graded paired comparison s usually ranges from 2 to 4. The larger the absolute value of $\mathbf{y}_t$, the stronger the preference for the associated item. The PCS logit model is specified by a generalized linear model with a logit link function: the stochastic component of the model is based on the preference $\mathbf{y}_t$ suitably recoded on a 0:1 scale, while the systematic component is based on a design matrix $\mathbf{X}$ with P columns describing the PCS tasks:

$$E\left(f(\mathbf{y}_t)\right) = \boldsymbol{\mu}, \quad \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}, \quad logit(\boldsymbol{\mu}) = \boldsymbol{\eta}\,. \tag{1}$$

The P-dimensional vector of parameters $\boldsymbol{\beta}$ s represent the items' utilities and they can be estimated by the maximum likelihood method. This approach is particularly indicated for the graded PCS model, as it can model the strength of competition within each set. The analysis is usually carried out for the full sample or for major groups of respondents. However, given a large enough number of tasks with respect to the number of items to be assessed, individual-level analysis can be carried out. McCullagh and Nelder (1989) have provided exhaustive information about the estimation algorithm and asymptotic properties of the parameter estimates.

Another popular approach to estimate individual-level PCS scores is hierarchical Bayes (HB) analysis. HB is particularly indicated to estimate PCS individual utilities given only a few tasks assessed by respondents. This is accomplished by borrowing information from population information describing the preferences of other respondents. HB models estimate preference coefficients for a given respondent based on his or her responses as well as on responses of similar respondents. Consequently, more information is used in estimating individual utilities, thus it is possible to estimate a larger number of parameters or the same number with greater

precision than other approaches allow. HB estimates tend to be robust to mistakes or inappropriate answers due, for example, to tiredness. HB approach was first adopted for conjoint analysis where, as for PCS analysis, usually there are many heterogeneous units of analysis (the respondents) but for each unit only little information is available (tasks evaluations).

While until a decade ago researchers could only run basic analysis on PCS data allowing only aggregate-level logit estimation for studies investigating many items, nowadays software packages offer comprehensive analytical capabilities, and HB is probably the typical choice for PCS utility estimation as it allows individual-level analysis.

2. THE NEED FOR MORE INFORMATION

In recent years, PCS has gained a significant increase in popularity among market researchers, due to its potentiality and design and analysis simplicity. Currently, it is a widely adopted model in many research areas including automotive, fast-moving consumer goods (FMCG), healthcare, transport economics, etc. The launch and diffusion of commercial software for the analysis of PCS data has surely contributed to the recent success of this approach, by increasing user accessibility and thus making it available to non-statisticians. Sawtooth Software *MaxDiff* (Sawtooth Software, 2007) is probably the most popular package to analyse PCS data (due to their cross-selling strategy – most of their customers approach Sawtooth Software for their wide conjoint offering), however the software seems to handle only the short model. A more complete package is *Demia R-sw Tradeoff* (Demia Studio Associato, 2013) which has been specifically designed to handle both the short and the graded PCS models. Both packages support HB analysis for PCS data and they also handle MDS analysis.

With growing popularity more and more researchers need to better understand the potentialities and limitations of PCS, especially considering that PCS results are often not just presented to the final user in their raw form, but they might be used for additional statistical analyses, such as feeding a segmentation model (Dillon *et al.*, 1993).

To date, it is not very clear the role played by the different PCS exercise elements with respect to the results accuracy. For example, researchers often wonder to what extent using multiple design versions is beneficial and what scale is the most appropriate to collect the PCS assessments from the respondents. There are several elements to be considered in a PCS study, and all play a potentially key

role in the accuracy of PCS results, although their role has not been be properly quantified yet:

- number of items considered in the exercise;
- short model versus graded model;
- choice of the scale for the graded model;
- inclusion of a neutral / indifference point;
- number of times each item is presented to each respondent;
- number of PCS tasks in the questionnaire;
- type and number of design versions (blocks);
- number of respondents;
- preference homogeneity among respondents;
- preference homogeneity among items.

There is very little work done in this area, as most of the PCS literature focus on alternative analysis algorithms, on comparisons against other popular approaches such as ratings, rankings or MDS, or on practical applications of the approach. Corradetti and Furlan (2006) carried out an analysis of the impact of the number of PCS tasks on the quality of the results. In their work, they considered 12 items assessed through a graded PCS model with a 7-point scale inclusive of an indifference point. They varied the number of tasks from 1 to 8 and they found out that there is a linear loss of quality as the number of tasks decreases, and no evident threshold could be identified.

With our work, we intended to explore the impact of some of the above elements to provide some actionable insight for researchers involved in designing PCS exercises. In particular, we had two objectives. The first one was related to the common practice among researchers to create many design versions (blocks), as it is commonly believed that the more versions are assessed by respondents, the higher is the results accuracy. However, it is not clear if having more than one version actually helps, and, if so, how many versions should be ideally prepared. The other objective was related to the impact on accuracy of using a short model versus a graded model and, in the latter case, the impact of the scale.

In order to meet these objectives, we decided to base our analysis on 15 items as, based on our experience, most PCS projects require the analysis of 12 to 18 items. We created a number of design combinations by varying the following three elements:

- PCS model: short, graded with 5 points, graded with 7 points. The indifference point had always been available (see Figure 1);

- number of tasks to be assessed by each respondent: 12, 15, 18;
- number of design versions (blocks) to be assigned to the respondents sample: 1, 2, 5, and 10.

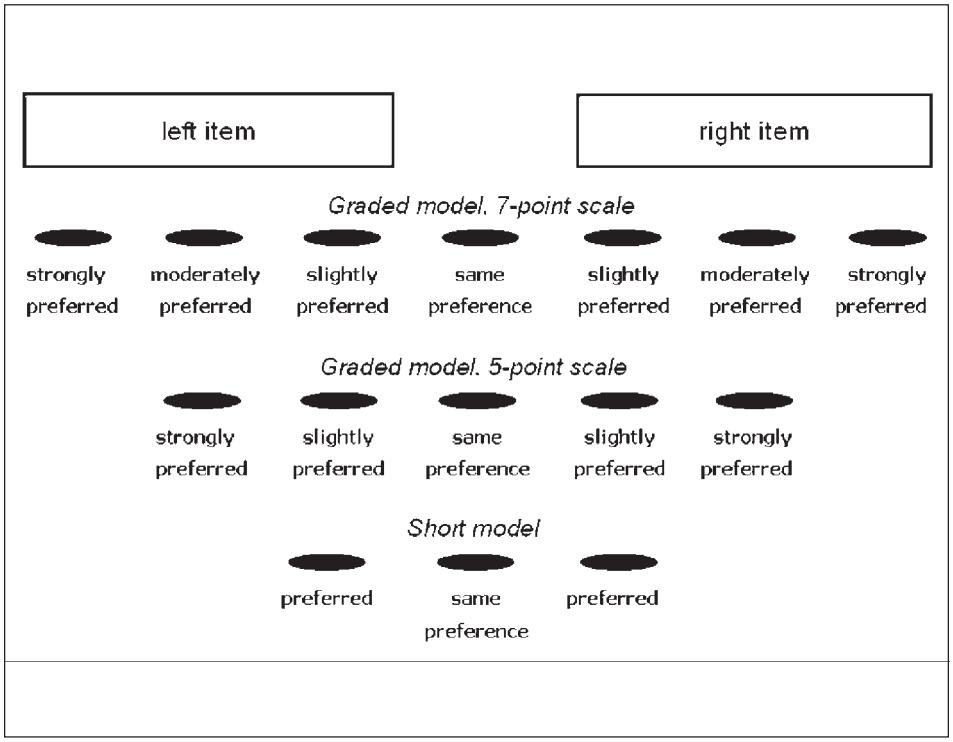

Figure 1: PCS models tested in this paper.

Table 1 provides an example of one of these designs. Three elements (PCS model, number of tasks, and number of design versions) characterise a 3x3x4 full factorial design (36 combinations).

In addition to the 12, 15, or 18 tasks assessed by the respondents, we also included some holdout tasks for validation purposes. Usually, just a couple of holdout tasks are included in a PCS exercise to keep it manageable and keep additional fatigue to a minimum. However, as reported in the next section, we used simulated respondents, thus we could include far more validation tasks (15 were presented to each respondent).

Table 1 – Design with 2 versions and 12 tasks. The 15 items appear as left or right elements across the various tasks

Version	Task	Left Element	Right Element
1	1	2	1
1	2	12	6
1	3	3	7
1	4	11	10
1	5	15	8
1	6	13	11
1	7	4	9
1	8	4	6
1	9	3	14
1	10	10	12
1	11	5	3
1	12	7	9
2	1	14	2
2	2	11	2
2	3	9	3
2	4	13	5
2	5	6	1
2	6	8	4
2	7	10	13
2	8	12	13
2	9	5	15
2	10	15	4
2	11	1	7
2	12	8	14

3. THE SIMULATION

In order to assess the impact of the PCS model, the number of tasks, and the number of design versions, we could not use results from real surveys, as they would inevitably be based on only one specific combination of these elements. Theoretically speaking, we could have administered alternative designs to the same sample, but this would not have been practical and we might also have risked introducing bias due to the fact that the same respondents would have already been exposed to a similar exercise a number of times. The only practical solution was simulating data, which consisted of two steps:

(1) simulating respondents' preferences (true utilities) for each statement;
(2) exposing these 'computerised' respondents to the alternative PCS design versions to obtain PCS data. These respondents would 'choose' the most preferred item (and expressing the strength of the preference in the graded model) from each pair according to their preferences simulated at (1).

In order to simulate the respondents' preferences, we looked at a number of previous PCS studies analysed with a HB model in order to assess what could be a reasonable distribution for each item. We noticed that the average of PCS preferences across the sample tends to be between 10 and 90 for most items (considering a scale 0:100). The distribution of these preference scores is asymmetric except, as one would expect, for items with an average around 50, with the asymmetry being the largest for the items whose average is closer to 10 (positive skewness) or closer to 90 (negative skewness). We fitted a beta distribution to each PCS item in every available project to assess potential beta coefficients for the PCS preference scores.

Based on this analysis of past studies, we generated preference scores for 15 items and 200 simulated respondents through a two-stage process. We chose this specific sample size as, based on our PCS projects review, this appeared the most common one, a good compromise between robustness and affordability. First, we randomly assigned an average preference score to each item between 10 and 90. Second, we generated scores for every item for each respondent by the addition of a beta distributed random variable with appropriate coefficients. The resulting generated scores were asymmetric with their distributions mirroring those seen in previous PCS studies.

As a second step, we had to give the various PCS design (i.e., 12 to 18 tasks) as well as the 15 holdout tasks to this set of 'computerised' respondents. For each task and respondent, the preference scores associated to the two items within a task were identified and thus transformed into an expressed choice based on an algorithm assumed to closely mirror the choice behaviour in the real world. This algorithm is based on the difference in preference score between the two items in the task and its structure depends on the model considered (short versus graded), as shown in Figure 2. The resulting choice data have the correct format to be analysed by the package R-sw Tradeoff (Demia Studio Associato, 2013) without further recoding.

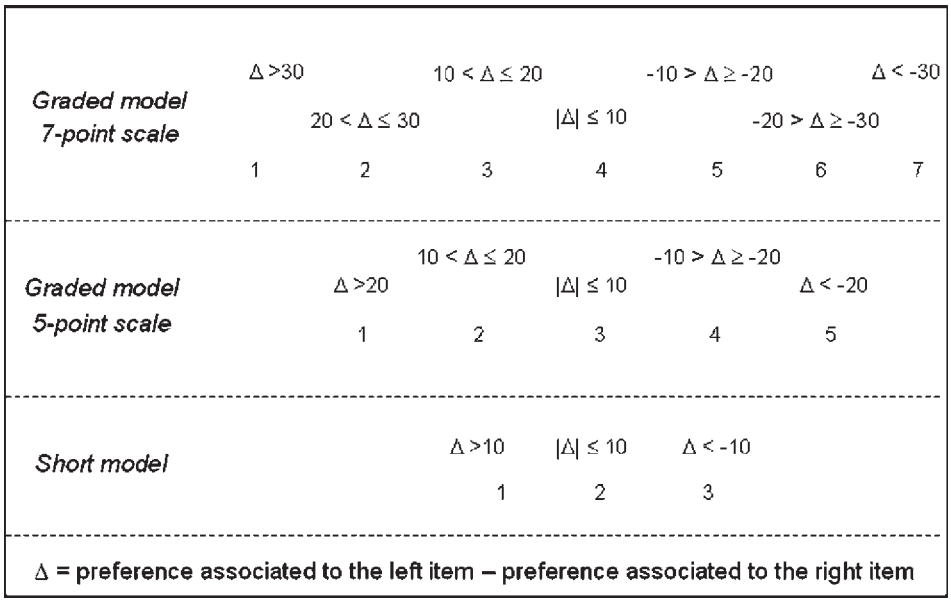

Figure 2: Rationale to convert preference scores into respondents choices

It is worth noting that no matter how well constructed are the 'computerised' respondents, a simulation will not be able to fully mirror the choice behaviour in the real world. In fact, in any trade-off exercise there is a certain amount of response error that might lead, for instance, to an item with higher utility not being preferred to an item with lower utility. However, although this limitation exists, we are confident that the simulation mirrors sufficiently well the actual choice behaviour in terms of order, context, and layout effects. These can be largely reduced and sometimes completely eliminated (e.g., when there are no prohibited combinations) through an accurate experimental design with an excellent one-way, two-way, positional, and within-block balance. One element, however, that could have an effect on the realism of the simulation is the number of tasks seen by each respondent and thus the length of the exercise. This is an effect that has not been well studied in relation to PCS projects and there appears to be the opportunity for further research. However, there is some evidence available for other trade-off models (e.g., conjoint analysis and discrete choice modelling) to suggest that this effect is hardly controllable, as it depends on many elements such as the target respondents, the complexity of the task, the respondent's level of engagement, etc. For this reason, we have decided not to introduce any adjustment coefficients in the simulation.

The design creation, the preference scores generation, the subsequent identification of the preference scores associated to the various items, the choice of the most preferred item, as well as the analyses described in the next section, were performed 100 times for each design combination in order to allow obtaining confidence intervals for the various outcomes of our analysis. The 95% asymmetric confidence intervals presented here are non-parametric percentile-based and were obtained by identifying the values of our statistics corresponding to the lower and upper 2.5% of the empirical distribution (Efron, 1981). Figure 3 illustrates the key steps of the process.

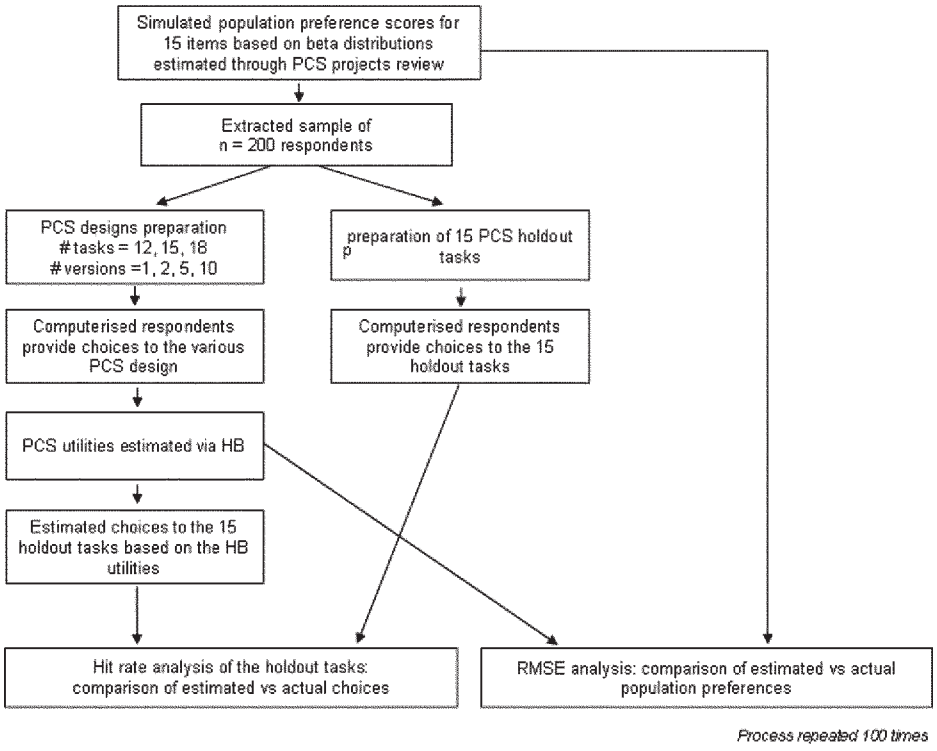

Figure 3: Key steps of the simulation process

4. DATA ANALYSIS AND VALIDATION

Once the above simulated datasets had been obtained, we could proceed with the analysis. We decided to analyse the data using HB modelling as this is the typical choice to obtain individual-level results; this approach works well even when there are many items to be estimated with respect to the number of tasks assessed by each

respondent.

Only the data associated to the PCS designs have been considered for the analysis, and not also the holdout tasks, which were considered later on for validation purposes only. For the HB analysis, we chose the *estimate.PCS.HB* function available in the package R-sw Tradeoff (Demia Studio Associato, 2013). This choice was dictated by the fact that this is the only commercial software that we are aware of that has been specifically designed to handle both the short and the graded models. This is a very flexible and convenient package; it has been possible to prepare an appropriate script to analyse all simulated datasets without repetitive and tedious manual intervention from us. For each alternative design and each respondents set we obtained a full set of PCS individual scores or utilities (a score for each item and respondent).

As a final but important step, we had to choose and adopt an appropriate approach to validate the quality of these sets of utilities against the original simulated preference scores. Thanks to our simulation framework, we had the assessment of 15 holdout tasks for each respondent, thus we could use a hit rate approach. We can say there is a hit when the PCS utility associated to the preferred item in the holdout task is larger than the PCS utility associated to the other item appearing on the task. Therefore, we defined as hit rate the percent of times that the HB model 'guesses' the preferred item. This analysis is based on all sample respondents and all holdout tasks they have been exposed to for which a preference was given either to the left or to the right item.

It is important to mention that, for the sake of an appropriate interpretation of results, the hit rate score for the model under investigation (i.e., the one based on the PCS individual scores) should be compared against the hit rate score of a random model (i.e., a model based on absolute randomness of choices, which is obtained by the reciprocal of the number of items presented in the various tasks). If the hit rate for the model under investigation is significantly higher than the one for a random model (in our case 1/2=50%), then it is possible to say that the model is, to some extent, satisfactory. The hit rate score of a random model represents a lower limit and is used to put the hit rate score into context.

From each design combination, in addition to the hit rate index, we computed the root mean square error (RMSE) between the population means on which we based our simulation (a set of 15 means varying between 10 and 90) and the HB utilities averaged across the respondents of the relevant sample (a set of 15 average scores). The RMSE has been important to assess how close the average HB utilities are to the true population values. While the hit rate index is based on the extracted samples and they are used to compare the prediction accuracy and thus the observed

versus the estimated ranks of the items, the RMSE indexes are based on the population characteristics and are used to compare the actual magnitude of the preference scores, not the ranks. Both metrics have been included in the analysis because the assessment of the impact of the different design elements on the results should take into account both the magnitude and the rank of the preference scores.

The hit rate and the RMSE indexes for each design combination are presented in Figure 4 and Figure 5. These figures also show the 95% percentile-based confidence intervals associated to each outcome based on the 100 iterations.

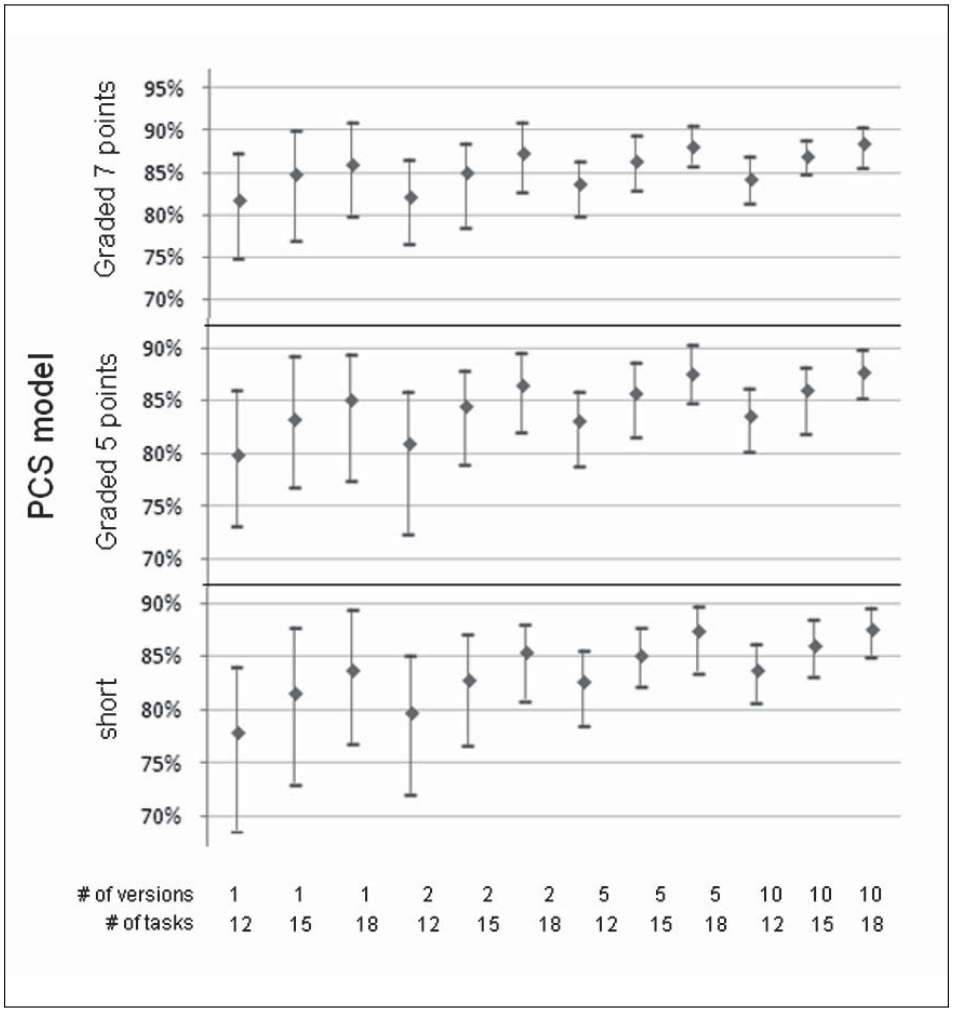

Figure 4: Effects of each design combination based on hit rates

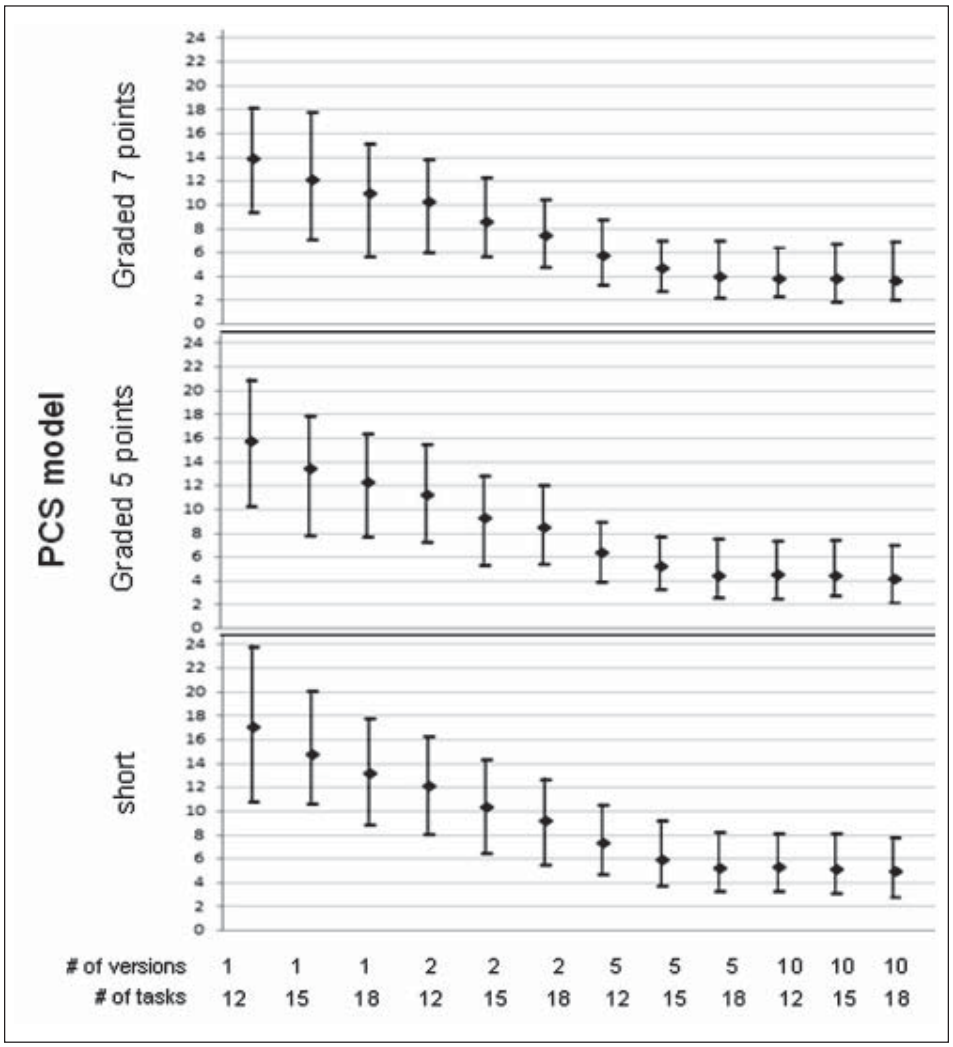

Figure 5: Effects of each design combination based on RMSE indexes

Figure 4 and Figure 5 show a clear pattern between the design elements and the metrics considered. In order to summarise the findings and separate the effect of each design element on the accuracy of the results, and thus obtain actionable information, we simply averaged the hit rate and the RMSE indexes across the various analyses that involved each design element. For example, to assess the effect of the 12 tasks, we averaged the hit rate and the RMSE indexes across all analyses that were based on 12 tasks. Summary results are presented in Figure 6 and Figure 7.

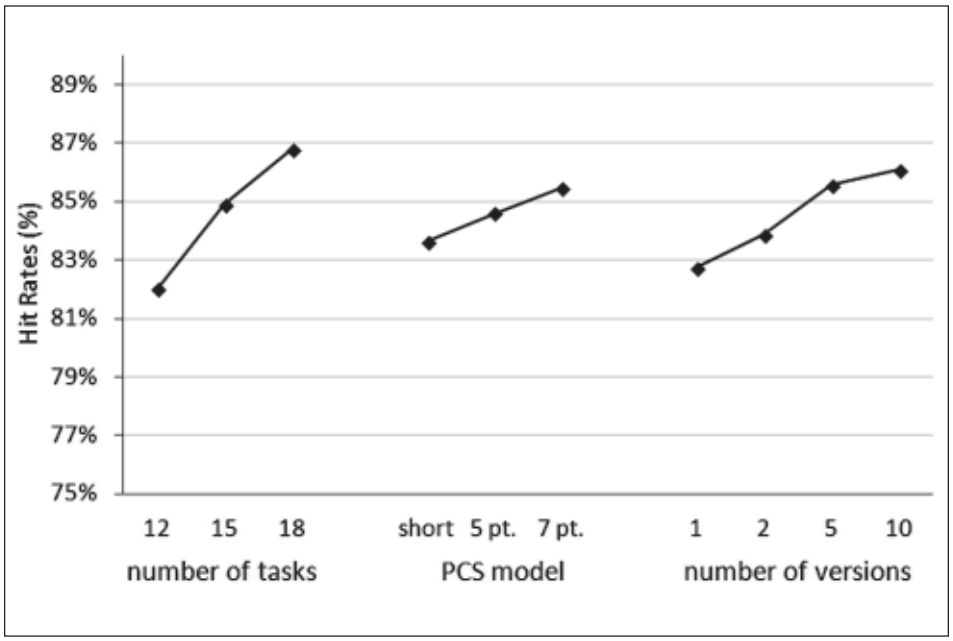

Figure 6: Summary of the effects of each design element based on hit rates

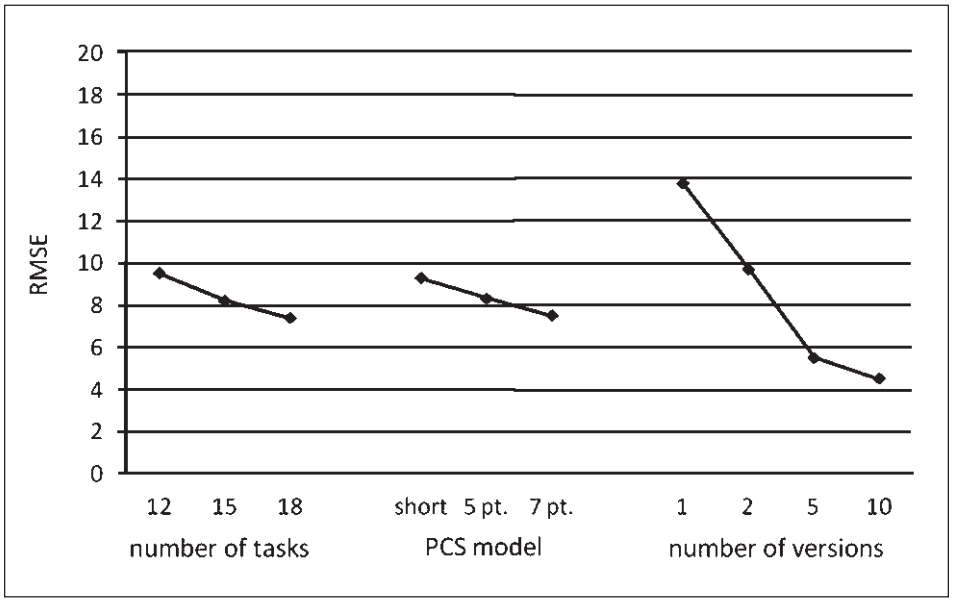

Figure 7: Summary of the effects of each design element based on RMSE indexes

Both Figure 6 and Figure 7 show that the three design elements we considered in our simulation project all have a significant impact on the results. The magnitude of the impact of the number of PCS tasks was expected as this is usually the main element, along with the number of items to be assessed, considered by the statistician when designing a PCS exercise. This impact seems to be almost linear and this is consistent with previous findings (Corradetti and Furlan, 2006).

On the contrary, the impact of the PCS model, although evident, was lower than we anticipated. It would appear that rather than adopting a more complex model, which might lead to respondent confusion and fatigue and thus lower data quality, especially among some target respondents, one could achieve the same accuracy by opting for a slightly longer exercise (more tasks) or by including a larger number of versions in the design. Therefore, while a more complex model increases the accuracy, it could be seen as a suboptimal solution.

The most interesting finding of our analysis is probably the impact of the number of design versions assigned to the respondents sample on results accuracy. With multiple versions of the questionnaire, different respondents see different series of tasks. It appears that the common belief that the more versions assessed by respondents, the higher the results accuracy holds true, with a distinct effect up to 5 versions. Having only one or few design versions does not provide any evident benefit to most projects, while it has a clearly detrimental impact on the accuracy. From our simulation, it appears that increasing the number of versions from 1 to 5 has an effect on results (based on hit rates) comparable to that of an increase in the number of tasks in the questionnaire from 12 to 15 and a slightly larger effect than using a more complex model (7-point graded model instead of the short model). It also appears that having more than 5 versions provides a further and evident benefit in terms of accuracy. While these results are truly impressive, it is worth noting that they probably underestimate the real benefits of multiple versions which are likely to be larger. Multiple versions considerably increase the variation in the way items are combined within tasks, which is likely to reduce potential context bias and order effects, both effects that have a psychological nature and cannot be assessed through a simulation.

The positive effects of multiple versions is extremely valuable for researchers, as increasing the number of versions is statistically easy to accomplish, economical, and most of the time practical (only with PAPI studies, nowadays no longer popular except in emerging markets, does having multiple versions prove challenging from a logistical point of view).

5. CONCLUSIONS

With this study we explored how some key elements in a PCS design impact on the accuracy of results. These results are consistent with, and complement well, previous research conducted in this area. Among the various potential elements we could focus on, we chose the type of PCS model, the number of tasks, and the number of design versions. We showed that it is preferable having a large number of design versions, not just one or two, although this might present logistics benefits under some isolated circumstances. Administering at least 5 design versions seems to considerably improve accuracy of results from a statistical point of view and this is also likely to reduce potential context bias and order effects which might have a negative effect on the quality of responses. We also showed that increasing the number of tasks has an important effect on accuracy, thus the researcher should include as many tasks as practical in the PCS exercise, but not so many to introduce elements of fatigue and confusion among respondents. Finally, we showed that also the type of the PCS model adopted for the exercise has an effect on accuracy, with more complex evaluation frameworks providing more accurate results. However, this effect, although evident, was not as strong as we expected. While a more complex evaluation framework in theory allows achieving a higher accuracy, it might increase respondent confusion and fatigue and thus lower the overall data quality, wiping out any positive effect. This represents a valuable finding for researchers who are recommended to keep the task as simple as possible, at least when the exercise is administered to non-experienced respondents.

It is important to highlight that the results obtained in this study are valid for a project with 15 items and 12 to 18 tasks included in the questionnaire, and they might be slightly different with a different project setup. For instance, with a larger number of items, there might be a higher threshold in the number of versions required to optimise results accuracy. Moreover, our findings are valid only for the simulation model we adopted. Results could have been different if another model were appropriate, for instance if average preference scores differed by larger or smaller amounts, their variability was different, or if they followed a lognormal, a gamma, or some other distribution. Further research is needed to assess to what extent the results are affected by the simulation model.

Further research is also required to assess different project setups and to investigate elements that have not been considered in this or previous studies, such as preference homogeneity among respondents and among items. Some additional research is also required to assess the impact of these elements in relation to various respondent types (e.g., busy professionals, young or old respondents) and in different fields (FMCG, B2B, durables, etc).

REFERENCES

Chrzan, K. and Golovashkina, N. (2006). An empirical test of six stated importance measures. In *International Journal of Market Research*, 48(6): 717-740.

Cohen, S.H. and Neira, L. (2003). Measuring preference for product benefits across countries. Paper presented at the *Sawtooth Software Conference 2003*, San Antonio, TX.

Corradetti, R. and Furlan, R. (2006). A trade-off model for eliciting physicians' or patients' preferences in healthcare research projects. *Memorie della Accademia delle Scienze di Torino* – Classe di Scienze Fisiche, Matematiche e Naturali – Applicazioni della Matematica, Classe V, 30: 55-89.

David, H.A. (1988). *The method of paired comparisons (second edition).* Oxford University Press, New York.

Dean A. and Voss, D. (1999). *Design and analysis of experiments.* New York: Springer-Verlag.

Demia Studio Associato (2013). *R-sw tradeoff: a package for advanced trade-Off analysis*. http://www.demia.it. Last access 03/09/2013.

Department of Trade and Industry (2002). *Car servicing and repairs: Research into trade and consumer interest in a national 'good garage' scheme.* Great Britain Department of Trade and Industry, London.

Dillon, W. R., Kumar, A., Smith de Borrero, M. (1993). Capturing individual differences in paired comparisons: An extended BTL model incorporating descriptor variables. In *Journal of Marketing Research*, 30: 42-51.

Efron, B. (1981). Nonparametric estimates of standard error: the jackknife, the bootstrap, and other resampling methods. In *Biometrika*, 68: 589-599.

Finn, A., and Louviere, J. J. (1992). Determining the appropriate response to evidence of public concern: the case of food safety. In *Journal of Public Policy and Marketing*, 11: 12-25.

GfK (2010). *Conjoint designs and techniques for healthcare.* http://www.gfknop.com/imperia/md/content/gfk_nop/healthcare/conjoint_designs_and_techniques_for_healthcare.pdf. Last access: 03/09/2013

Green, P. E., Srinivasan, V. (1978). Conjoint analysis in consumer research: issue and outlook. In *Journal of Consumer Research*, 5: 103-123.

Jackson, J. E. and Shenker, K. L. (1981). Incomplete paired comparisons for conjoint analysis. In *Proceedings of the Business and Economics Statistics Section*: 37–40.

John J. A., Wolock F. W., and David H. A. (1972). Cyclical designs. *Applied Mathematics Series*, 62. National Bureau of Standards, Washington, D.C.

Jaeger, S. R., Jorgensen, A. S. , Aaslyng, M. D., Bredie, W. L. P. (2008). Best-worst scaling: an introduction and initial comparison with monadic rating for preference elicitation with food products. In *Food Quality and Preference*, 19: 579-588.

Louviere, J.J. and Flynn, T. (2011). Advances in best-worst scaling (BWS) and choice-based measurement methods. Paper presented at the *AERE 2011 Pre-Conference Workshop*: *Recent Developments in the Design and Implementation of Discrete Choice Experiments*, Seattle, Washington.

Louviere, J. J. and Woodworth, G.G. (1990). *Best-worst scaling: A model for largest difference judgments*. Working paper, Faculty of Business, University of Alberta.

Madansky, A. (2010). Data Use: Evaluating paired comparisons, maximum difference and traditional ranking. In *Quirk's Marketing Research Review*, 24(10): 22-27.

McCullagh, P., and Nelder, J. A. (1989). *Generalized linear models*. London: Chapman and Hall.

Sawtooth Software (2007). The maxdiff/web system technical paper. *Sawtooth Software Technical Paper Series*, http://www.sawtoothsoftware.com/download/ techpap/maxdifftech.pdf. Last access: 03/09/2013.

Srinivasan, V. S. and Wyner, G. A. (2009). An improved method for the quantitative assessment of customer priorities. Stanford University Graduate School of Business, Research Paper No. 2028.

Steenkamp, J. B. E. M. and Hofstede, F. T. (2002). International market segmentation: issues and outlook. In *International Journal of Research in Marketing*, 19: 185-213.

Thurstone, L. L. (1927). A law of comparative judgement. In *Psychological Review*, 34: 273–286.