Transportation **21**: 167–184, 1994

Use of the logit scaling approach to test for rank-order and fatigue effects in stated preference data

MARK BRADLEY & ANDREW DALY
Hague Consulting Group, Surinamestraat 4, 2585 GJ Den Haag, The Netherlands

Received 11 January 1993; accepted 2 June 1993

Key words: discrete choice analysis, logit analysis, modelling, ranked data, stated preference, survey design

Abstract. The scaling approach is a statistical estimation method which allows for differences in the amount of unexplained variation in different types of data which can then be used together in analysis. In recent years, this approach has been tested and recommended in the context of combining Stated Preference and Revealed Preference data. The paper provides a description of the approach and a historical overview. The scaling approach can also be used to identify systematic differences in the variance of choices within a single Stated Preference data set due to the way in which the hypothetical choice situations are presented or the responses are obtained. The paper presents the results of two case studies – one looking at rank order effect and the other at fatigue effect. Scale effects appear to exist in both cases: the amount of unexplained variance is shown to increase as rankings become lower, and as the number of pairwise choices completed becomes greater. The implications of these findings for the use of SP ranking tasks and repeated pairwise choice tasks are discussed.

Introduction

Discrete choice analysis using *revealed preference* (RP) data based on actual travel choices has become the dominant estimation approach in disaggregate travel demand modelling. A comprehensive treatment of this approach is given by Ben-Akiva and Lerman (1985). Since the early 1980's, increasing use has been made of *stated preference* (SP) data based on stated choices under hypothetical contexts. With the use of controlled survey contexts, SP methods have proven very effective in estimating the relative importance of variables which influence travel behaviour.

Models based on SP data may not be appropriate for *predicting* behaviour if the amount of residual unexplained variation in the SP choice data differs from the amount of residual variation in actual (RP) choices. A number of authors have discussed the reasons why one might expect differing amounts of unexplained variance in choices from RP and SP survey contexts (Bonsall 1985; Bates 1988; Morikawa et al. 1990; Daly and Ortúzar 1990; Bradley

and Kroes 1992). Because SP experiments can be isolated to some extent from influences outside the survey context, one usually expects less unexplained variance in SP choice data. On the other hand, the survey instrument itself can introduce influences on choice which are not present in actual choice situations. To make SP models more applicable to predicting actual choices, there has been a growing emphasis on customising SP choice contexts as closely as possible around actual choice situations recently faced by respondents.

With the objective of combining the best aspects of RP and SP data, it is desirable to use them together in analysis. If possible differences in residual variance in the data are not accounted for, however, such analysis may give biased results. The *scaling approach* is a statistical estimation method which allows for differences in the amount of unexplained variation in different types of data which are used together in analysis. This approach is introduced in further detail in the following section. The results of two case studies are described in the subsequent sections, and a brief summary is given in the final section.

The Scaling Approach

The objective of discrete choice statistical estimation techniques such as logit and probit is to estimate a utility function which best predicts the choices in an observed sample. Supposing that we have two types of data, the utility functions to be estimated may appear as:

$$U_1 = \beta.X_1 + \alpha.Y + \mu_1$$

$$U_2 = \beta.X_2 + \Gamma.Z + \mu_2$$

where:

- β is a vector of parameters to be estimated, assumed to have the same values in both data sets;
- X_1, X_2 are vectors of observed values of variables common to the two data sets;
- Y, Z are vectors of observed variables which may be specific to one data set or the other;
- α, Γ are vectors of parameters to be estimated for the data-specific inde pendent variables;
- μ_1, μ_2 represent the amount of residual, unexplained variance in the choices in the two data sets.

The correct values of β are assumed to be identical for both data sets. In

practice, this assumption is not too restrictive, given that the data-specific variables in Y and Z can be used for effects which apply only to certain types of data. The random error terms, μ_1 and μ_2 are assumed to incorporate all unobserved or unspecified effects on the choices. These terms are assumed to be independently and identically distributed across all observations within a data set. Since "utility" has no absolute scale, it is not possible to estimate optimal values for both the explanatory parameters *and* the residual error term. In typical estimation procedures, the distribution of the residual error term is assumed in advance (standard normal for probit, Gumbel for logit) and the scale of the coefficients is estimated *relative* to this distribution.

Suppose we were to combine two types of data in a single model assuming a single random error variance. If, in reality, one type of data systematically has more unexplained variance than the other, this can lead to two problems:

- *The model parameters may be biased:* The data type with systematically larger random variance should receive less "weight" in estimating the β coefficients for the observed variables. If this is not the case, the estimated coefficients may be biased.
- *The model elasticities may be too high or too low:* In application, the sensitivity of the choice predictions to changes in the X variables is a function of the overall scale of the β parameters relative to the variance of error term μ. If, therefore, observations are used in estimation which have a variance much different from that in the context for which predictions are made, the model will tend to over- or underpredict actual changes in choice probabilities.

In the notation above, such problems could occur if we attempt to estimate the utility functions U_1 and U_2 in a single model which assumes constant variance across observations when, in reality, μ_1 and μ_2 do not have the same distribution. The scaling approach addresses this problem by allowing different types of data to have different error variances within a single model. Suppose that:

$$\theta^2 = variance(\mu_1)/variance(\mu_2)$$

Scaling the utility of data set 2:

$$U'_2 = \theta.U_2$$

we thus allow U'_2 to be estimated in a single model with U_1 in an efficient and unbiased manner. With the multiplicative scale parameter θ, however, the utility function U'_2 is no longer linear in the parameters, and standard model estimation methods are no longer appropriate. The main challenge has thus been to develop a feasible and accessible model estimation procedure to

estimate the scale factor(s) θ at the same time as the other unknown parameters (β, α, Γ).

Ben-Akiva and Morikawa (1990) applied a procedure to efficiently estimate the scale differences between different RP and SP data sources. This procedure maximises a joint likelihood function for observations from two or more data sources used simultaneously. As long as the utility function for each data source has at least two β parameters in common with those for the other data sources, a relative scale factor θ can be estimated for each type of data (except one which is arbitrarily chosen as the base data type with scale of 1.0). This maximum likelihood procedure was specially programmed in the GAUSS language and applied to a data set from the Netherlands which contained RP observations as well as observations from two separate SP experiments. The results are reported in Morikawa (1989).

Bradley and Daly (1992) incorporated the one-step estimation approach of Morikawa and Ben-Akiva into one which can be carried out with any logit estimation software capable of estimating models with nested "tree" structures (Daly 1987). They present two case studies, the first of which used the same Dutch data set us used by Morikawa and yielded nearly identical results. The second case study was based on much larger RP and SP data sets from Australia. This was a more rigorous test, given that the RP and SP data sets were collected from different samples and survey instruments. The Australian case study also demonstrated the ability of the logit-based scaling approach to handle multinomial and nested (tree) structures within both the SP and RP utility functions – in this case a nested choice from among five long-distance travel modes. The scaling approach is implemented using a "artificial tree" structure to take advantage of existing software capabilities. More details can be found in the Bradley and Daly (1992) paper, in Hensher and Bradley (1993), and in the paper by Hensher included in this issue.

The scaling approach may be useful in other contexts besides the mixed analysis of RP and SP data. Ben-Akiva and Morikawa (1990) note other possible contexts, such as the mixture of data from different types of RP data – e.g. household surveys, roadside interviews, traffic counts, etc. Swait and Louviere (1993) use a sequential scaling approach to combine data from an SP choice task (choice of A vs. B) and an SP rating task (A is acceptable vs. unacceptable), both based on the same set of variables. They found that the relative importance of the variables estimated from the two types of response data was the same, but that the choice task gave somewhat more precise information (less residual variance) than the rating task.

The sources of variance *within* a single SP data set can also be investigated using the scaling approach. An often-discussed feature of SP data is the repeated measures aspect – the complex mixture of within-person and between-person variance in the responses. Unfortunately, this aspect is not

easily addressed using the scaling approach because there is often no way of separating specific types of observations as different "data sets". The extreme option of treating each person in the sample as a separate data set with its own scale factor is not practical. A promising approach for dealing with some aspects of the SP repeated measures issue may be models which account for serial correlation, such as those used in analysing longitudinal data (Morikawa et al. 1992).

Certain aspects of the repeated measures problem *can* be addressed, however, if we can account for that part of the within person variance which arises from the survey instrument itself. One could imagine, for example, that the survey method could introduce differences in variance between observations from a single choice experiment. For a ranking exercise, people may pay more attention (less "random" error) to the ordering of the best options that they do to the ordering of the worst options. For a series of pairwise choice exercises, respondent "fatigue" may cause people to make choices less carefully as the number of choices increases.

The case studies described in the following two sections are meant to illustrate the use of the scaling approach to SP data in a context other than mixed SP/RP analysis. In the process, we investigate the SP rank order and fatigue effects just hypothesised. The questions which we address are: (1) do such effects seem to exist and (2) if so, what effect do they have on the estimation results?

Case Study 1: Rank order effects

The first case study is based on SP data collected during a study conducted for Stockholm Transport, described in Widlert et al. (1989). The data was collected during home interviews with over 300 people who commuted by bus. Separate SP experiments included variables related to bus service levels, bus stop facilities and bus vehicle factors respectively. These three experiments were administered to each person in random order during the interviews.

Each experiment contained nine alternatives, created using standard fractional factorial designs. Two additional alternatives were created – one with the best possible levels of all variables, and the other with the worst possible levels of all variables. Cards with the two extreme options were placed at the ends of a metre stick. The respondent then ranked the other nine design option cards in preference order by arranging them along the metre stick between the best and worst options.

Analysis was done using the *exploded logit* method in which a ranking of N alternatives is treated as N-1 independent observations: rank 1 chosen over ranks 2 through N, rank 2 chosen over ranks 3 through N, . . . , rank N-1 chosen

over rank N. This analysis method has often been subject to criticism; although the required distributional assumptions for logit estimation are satisfied, it is questionable whether the utility assumptions underlying discrete choice methods such as logit are applicable to rank data (Daly 1990).

Estimation results for the three experiments are presented in Tables 1 to 3. The first ("Base") model in each table was estimated using the SP design variables in dummy (0/1) variable form. All three experiments had 3 levels of fare, allowing estimation of two fare variables. Each experiment also contained three other variables, each with either 2 or 3 levels. Each experiment was different in terms of the number of levels and the experimental design used. The designs are given in Appendix A.

In the "Base" models, all coefficients are significant with the expected signs, except for "automated ticketing only" in Table 2 which gave a somewhat ambiguous result, with some people for it and some against it. Note that the significance (*t*-statistics) for these models may be overstated because the repeated measures nature of the data has not been accounted for in any way.

For the second "Scaled" model in each table, the data was essentially split into eight separate data sets, one corresponding to each "chosen" rank in the order – all 1st choices as chosen, all 2nd choices as chosen, etc. The first

Table 1. Experiment 1 – service levels (302 respondents, 2416 observations).

Model	Base	Scaled	Simulated scaled
Log-likelihood (b)	–3161.2	–3073.2	–3144.0 (Simulated base = –3145.9)
Coefficients *(t-statistics w.r.t. 0)*			
Fare up by 20%	–0.952 (–15.5)	–1.425 (–8.9)	–0.831 (–9.6)
Fare down by 20%	0.550 (9.7)	1.241 (9.7)	0.629 (9.0)
More punctual	0.792 (14.0)	1.468 (9.8)	0.768 (10.0)
Less punctual	–0.939 (–15.3)	–1.720 (–10.7)	–0.948 (–9.7)
No interchange	0.561 (11.1)	1.523 (8.2)	0.554 (7.6)
Travel time –20%	0.531 (10.5)	0.656 (6.7)	0.578 (9.0)
Scale factors *(t-statistics w.r.t. 1)*			
Rank 1 (base)		1.000 (–.–)	1.000 (–.–)
Rank 2		0.725 (–3.3)	1.074 (0.7)
Rank 3		0.704 (–3.7)	1.157 (1.2)
Rank 4		0.792 (–2.1)	1.098 (0.8)
Rank 5		0.472 (–8.5)	1.009 (0.1)
Rank 6		0.217 (–19.1)	1.003 (0.0)
Rank 7		0.431 (–10.0)	0.939 (–0.5)
Rank 8		0.241 (–12.6)	1.032 (0.2)

Table 2. Experiment 1 – bus stop facilities (294 respondents, 2352 observations).

Model	Base	Scaled	Simulated scaled
Log-likelihood (b)	–2915.7	–2846.4	–2880.9 (Simulated base = –2883.6)
Coefficients *(t-statistics w.r.t. 0)*			
Fare up by 10%	–0.902 (–14.9)	–1.496 (–10.7)	–0.843 (–12.2)
Fare up by 20%	–1.861 (–26.6)	–3.436 (–12.3)	–1.740 (–15.4)
Bus shelter present	1.840 (29.2)	3.802 (11.1)	1.779 (14.1)
Real time inform.	1.131 (19.8)	1.693 (11.5)	1.034 (14.4)
Automated ticketing	–0.242 (–4.6)	–0.054 (–0.4)	–0.210 (–4.1)
Scale factors *(t-statistics w.r.t. 1)*			
Rank 1 (base)		1.000 (–.–)	1.000 (–.–)
Rank 2		0.631 (–5.8)	1.077 (0.9)
Rank 3		0.391 (–13.4)	1.125 (1.3)
Rank 4		0.493 (–9.4)	1.043 (0.4)
Rank 5		0.533 (–7.4)	1.053 (0.5)
Rank 6		0.317 (–12.1)	1.269 (1.6)
Rank 7		0.276 (–11.8)	0.981 (–0.2)
Rank 8		0.239 (–11.1)	1.006 (0.0)

Table 3. Experiment 3 – bus vehicle factors (300 respondents, 2400 observations).

Model	Base	Scaled	Simulated scaled
Log-likelihood (b)	–3021.6	–2904.4	–2995.0 (Simulated base = –3001.1)
Coefficients *(t-statistics w.r.t. 0)*			
Fare up by 20%	–0.584 (–10.1)	–2.150 (–7.6)	–0.528 (–7.9)
Fare up by 40%	–1.871 (–26.4)	–5.574 (–9.0)	–1.856 (–11.9)
No seat for 2 min.	–1.063 (–18.0)	–3.201 (–8.9)	–1.047 (–11.8)
No seat for 10 min.	–1.974 (28.5)	–5.748 (–9.7)	–1.995 (–12.4)
Clean inside	0.374 (6.4)	1.650 (5.8)	0.334 (5.0)
Clean inside + out	0.387 (6.5)	2.915 (6.0)	0.344 (4.2)
Destination signs	0.270 (5.4)	0.100 (0.6)	0.224 (4.4)
Scale factors *(t-statistics w.r.t. 1)*			
Rank 1 (base)		1.000 (–.–)	1.000 (–.–)
Rank 2		0.541 (–7.8)	1.043 (0.4)
Rank 3		0.430 (–11.0)	0.931 (–0.7)
Rank 4		0.457 (–8.7)	0.920 (–0.8)
Rank 5		0.281 (–17.1)	1.141 (1.1)
Rank 6		0.125 (–35.1)	1.074 (0.5)
Rank 7		0.190 (–26.9)	1.329 (1.8)
Rank 8		0.190 (–27.9)	0.912 (–0.5)

rank was set as the "base" data set, and the logit scaling approach was used to estimate seven scale factors for ranks 2 to 8 relative to rank 1. The model was specified in the ALOGIT package using an "artificial tree structure" to nest the alternatives corresponding to each rank order, as illustrated in Appendix B.

The scale factor estimates are shown in the tables along with the *t*-statistics relative to 1.0 (no difference in variance) All scale factors are significantly less than 1.0, indicating that the amount of unexplained variation increases with lower rankings. For all three models, the scale factors for ranks 2 to 4 are roughly in the range 0.50 to 0.75, while the scale factors for ranks 5 to 8 are roughly in the range 0.20 to 0.50. The decreasing trend can be seen clearly in Figure 1.

In terms of model fit, the addition of seven scale parameters in the "Scaled" models adds 88, 69 and 117 log-likelihood units for the three experiments with respect to the "Base" models. These are highly significant improvements according to the likelihood ratio test. The t-statistics of the design variables are generally reduced by one half to one third in the "Scaled" models relative to the "Base" models. One could consider these values to be closer to the "true" significance of the effects, since more aspects of behaviour have been accounted for. Note that the two smallest coefficients, for automated ticketing and destination signing, are no longer significant.

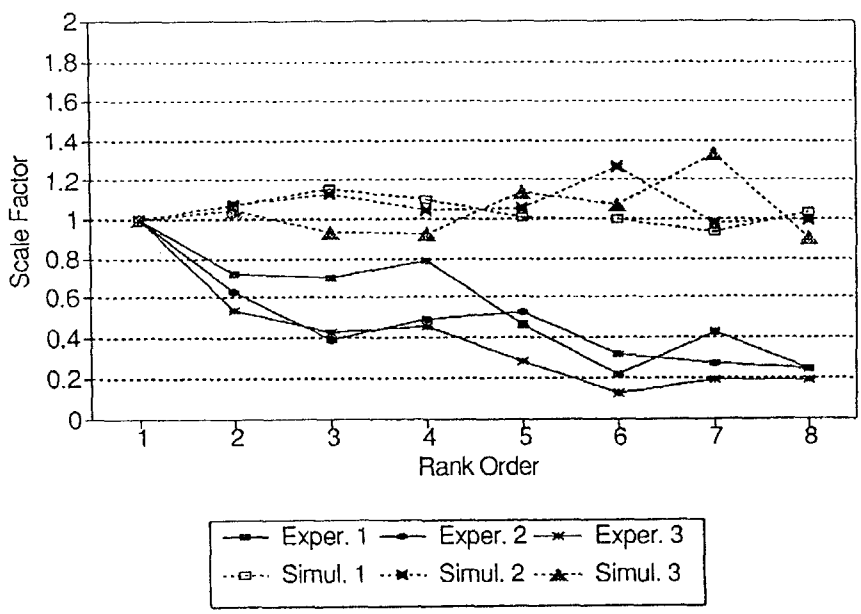


Fig. 1. Scale factors for exploded rank models.

Not only does the scale and significance of the coefficients change in the "Scaled" models, but the relative magnitudes sometimes shift as well. In Table 1, for instance, no interchange becomes more important relative to the other variables, while travel time savings becomes less important. In Table 3, cleanliness of the buses becomes more important relative to the other variables, while destination signing becomes less important. Grouping the data from different rankings without accounting for scale differences thus appears to have biased some of the results away from the values obtained when scale differences are accounted for.

An even stronger test than estimating scale factors is to estimate completely separate models for each of the eight rank "data sets" and then look at the change in total likelihood. Such tests (not shown in the tables) gave improvements of 175, 108 and 180 log-likelihood units with respect to the three "Scaled" models, for the addition of 36, 30 and 42 parameters, respectively. These are significant improvements, indicating that the scale differences along do not explain all of the differences between the rank orders. Ben-Akiva et al. (1991) also found significant parameter differences corresponding to rank order.

We were concerned that our results might be an artefact of the analysis method used. To test this, the estimated utility functions from the "Base" models were used to simulate rank-order responses, and these were substituted for the actual responses in the SP data sets. The simulated responses assumed no rank-order scale effects: the same error variance was assumed for every alternative (simulated error terms were drawn randomly from a standard Gumbel distribution).

The same "Base" and "Scaled" model specifications shown in Tables 1 to 3 were estimated using the simulated rankings for the three experiments. The "Simulated Base" model results (not shown) were essentially the same as the "Base" results for the actual rank data in terms of coefficients, *t*-statistics and likelihood. The "Simulated Scaled" model results are shown in the final column of each table. The estimates remain close to the "Base" model coefficients used to simulate the data, and, contrary to the models on the actual data, the rank-specific scale factors are not statistically different from 1.0. Figure 1 shows the contrast between the actual and simulated scale results more clearly. Furthermore, the improvements in log-likelihood over the "Simulated Base" models are only 2, 3 and 6 units, which are not significant for the addition of 7 scale parameters. Further tests estimating completely separate models for each simulated rank (not shown) gave improvements of 15, 11 and 22 log-likelihood units over the "Simulated Scaled" models – again not significant. The simulation results provide strong evidence that the scale effects are due to differences in the way in which real respondents make decisions at different points in the ranking process.

Our findings are corroborated by those from an earlier study by Hausman and Ruud (1987), who also identified increasing amounts of unexplained variance corresponding to lower rankings. These findings have negative implications for the use of rank-order SP data and exploded logit, which are discussed further in section 5.

Case Study 2: Fatigue effects.

The second case study is based on SP data collected for the Dutch Railways to study train/car mode choice for intercity travel in the Netherlands, described in Bradley et al. (1988). This same data has been used in the context of joint RP-SP analysis by Morikawa (1989) and Bradley and Daly (1992). Here, we use only the data from the within-mode SP experiment which offered choices between train service options.

The survey was administered using the MINT computer-assisted personal interview (CAPI) software. Four design variables were used, each with two or three levels. The levels are given in Appendix C. An orthogonal fractional factorial design of nine alternatives was randomly selected for each respondent from the full factorial design of all possible combinations. Respondents were presented with a series of pairwise choices from among these nine options. The first pair offered was a "dominant" choice, where one option was clearly superior to the other. This first pair served as a lead-in to the experiment, allowing the interviewer to check whether the respondent understood the choice task. From the second choice onwards, pairs of alternatives were presented in random order, until the point where the software could infer a full preference ordering across the nine alternatives, assuming transitivity. On average, each respondent completed about 13 pairwise choices, with all respondents completing between 10 and 16 choices.

Binary logit models were estimated on the pairwise choice data, using linear functions of the four design variables – train fare, travel time, number of interchanges and comfort level. Estimation results are presented in Table 4. Model 1 is the "Base" model, assuming no scale differences between observations. The coefficients are all significant with the expected sign.

The second column shows the "Scaled A" model. Here, the data was separated into 15 "data sets": all choices which were done 1st, all which were done 2nd, and so on. Because the first choice faced by each respondent was an obvious dominant comparison, the second choice was specified as the base "data set" with scale 1.0. Fourteen scale factors were estimated for the other response orders relative to the second response. The same four design variable coefficients were specified to apply to all responses. The "artificial tree" structure used in ALOGIT is shown in Appendix D.

Table 4. Train service pairwise choices (243 respondents, 2929 observations).

Model	1-Base	2-Scaled A	3-Scaled B
Log-likelihood	–1724.2	–1668.2	–1670.1
Coefficients (t-statistics w.r.t. 0)			
Train fare (fl)	–0.1484 (–19.9)	–0.1723 (–6.3)	–0.1900 (–9.2)
Travel time (min.)	–0.0287 (–10.7)	–0.0332 (–5.4)	–0.0356 (–6.6)
No. of transfers	–0.3263 (–5.5)	–0.3412 (–3.3)	–0.3884 (–3.9)
Comfort level	0.9457 (14.6)	1.1400 (5.8)	1.2500 (8.1)
Scale factors (t-statistics w.r.t. 1)			
Response 1		3.392 (2.8)	3.096 (3.2)
Response 2 (base)		1.000 (–.–)	1.000 (–.–)
Response 3		1.219 (0.8)	" "
Response 4		0.962 (–0.2)	0.764 (–1.9)
Response 5		0.735 (–1.5)	" "
Response 6		0.830 (–0.9)	0.709 (–2.4)
Response 7		0.729 (–1.5)	" "
Response 8		0.694 (–1.7)	0.657 (–2.9)
Response 9		0.752 (–1.3)	" "
Response 10		0.629 (–2.0)	0.565 (–3.6)
Response 11		0.616 (–2.3)	" "
Response 12		0.365 (–3.8)	0.332 (–5.7)
Response 13		0.355 (–3.2)	" "
Response 14		0.588 (–1.7)	0.363 (–4.7)
Response 15+		0.242 (–4.0)	" "

The resulting improvement in log-likelihood relative to the "Base" model is 56 units, which is highly significant for the addition of 14 parameters. The scale for the 1st response is quite high, as one would expect – there is little chance for error to affect such obvious choices. The scale for the 3rd response is somewhat higher than for the second, but from the 4th response onward the scale is consistently less than 1.0. Although the standard errors for the scale factor estimates are fairly high due to the limited sample size in each "data set", one can see in Figure 2 that a clear trend is present. A "fatigue" effect (higher unexplained variance) appears to set in around the time of the 5th response and to become much stronger by the time of the 12th response. Because the scale of the first choice appears to be quite different than the rest, the models were reestimated omitting the first observation per respondent. The results did not change noticeably from those in the table.

In the third model in Table 4 ("Scaled B") the response orders are grouped into pairs, so that only 7 scale factors are estimated. Compared to the "Scaled A" model, the loss in log-likelihood is not significant - only 2 units with 7

fewer parameters. With the larger sample sizes, the scale parameters are now significantly different from 1.0. The trend in the scale factors shows in Figure 2 appears virtually unchanged from that of the "Scaled A" model.

The *t*-statistics for the "Scaled B" model are one half to one third lower than those for the "Base" model – the same result was found in the first case study. In contrast to the first case study, however, the scaling approach here has almost no effect on the relative magnitude of the coefficients. The inferred monetary values of time for the three models in Table 4 are 11.60, 11.56, and 11.24 guilders per hour. Also in contrast to the rank-order case study, estimating completely separate models for each response order (not shown in the table) gave no significant improvement in total likelihood over the scaled models (33 log-likelihood units for the addition of 45 parameters). Possible reasons for the contrasts between the rank order and fatigue results are discussed in the final section.

We tested whether the results in Table 4 were due to the scaling approach itself. The estimated utility function from the "Base" model was used to simulate pairwise choice responses, which were substituted for the actual responses in the SP data. The "Base" and "Scaled B" model specifications in Table 4 were then estimated on the simulated choices. None of the estimated scale factors were significantly different from 1.0 (results not shown in the table.) Figure 2 shows the clear difference between the results for the actual

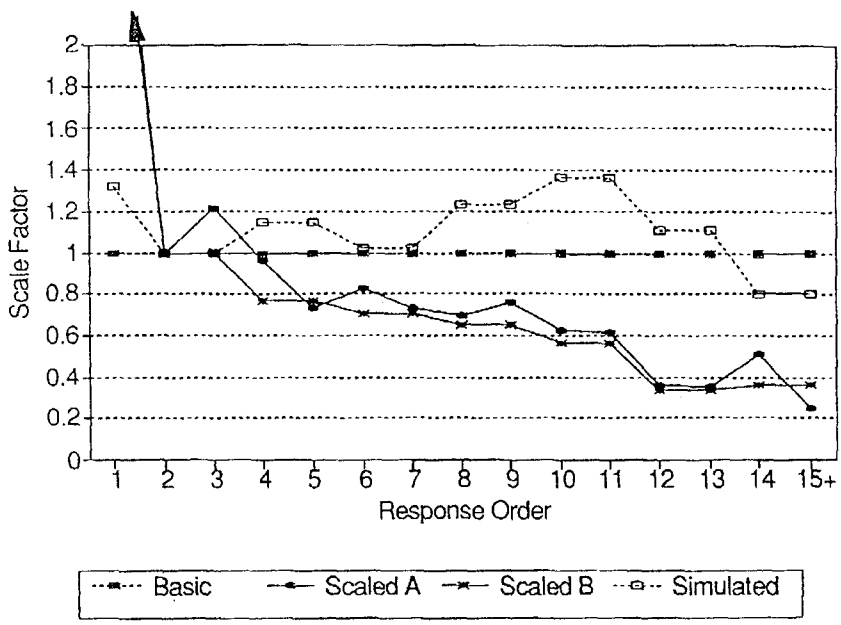


Fig. 2. Scale factor for fatigue effect models.

and simulated choices. For the simulated data, the "Scaled B" model increased the log-likelihood by only 4 units with respect to the "Base" specification, not significant for the addition of 7 scale parameters. Again, the simulation tests provide evidence that the "fatigue" effect is a real-world phenomenon.

Summary and conclusions

The logit scaling approach provides an efficient one-step estimation procedure to account for differences in the amount of unexplained variance when using different types of data together in estimation. The approach has usually been presented in recent literature in the context of combining RP and SP data. In this paper, we demonstrate the use of the logit scaling approach to account for survey effects *within* SP data from a single experiment. In two case studies, we have tested for the presence of a "rank-order effect" in rank data and of a "fatigue effect" in repeated pairwise choice data.

The results from the two case studies show a number of similarities. Scale effects appear to exist in both cases: the amount of unexplained variance is shown to increase as rankings become lower, and as the number of pairwise choices completed becomes greater. In both cases, the *t*-statistics of the design variable coefficients generally decrease by one half to one third compared to the base "naive" estimation results. The overall fit of the models, however, substantially improves due to the addition of the scale parameters. In both cases, these effects could not be reproduced using simulated response data, indication that the rank-order and fatigue effects are caused by the influence of the experimental tasks on real respondents.

We also obtained some contrasting results from the two case studies. For the pairwise choice data, the addition of the scale factors to account for respondent fatigue did not significantly change the relative magnitude of the model coefficients. For the rank-order data, on the other hand, some coefficients became more or less important relative to the others, and estimating entirely separate models for different positions in the rank order gave a significant improvement in likelihood relative to using only scale factors. Similar results were obtained by Ben-Akiva et al. (1991).

The pairwise choice data used in the second case study was collected using a computer-based approach which presented the pairs of alternatives in a different random order for each respondent. As a result, any adverse influence of the order-related fatigue effect may have been "randomised out" of the data. If a survey approach had been used where everyone received the same pairs of alternatives in the same order, the task order could be correlated with the design levels, and thus the scale differences could influence the relative parameter estimates. For SP ranking exercises, where the rank order is deter-

mined by the respondent, it would be difficult to eliminate any relationship between the rank order and the levels of the design variables. Some of the effects may be eliminated, however, by using block experimental designs which present different groups of respondents with different sets of alternatives for ranking.

For pairwise choice data, the results indicate that strong fatigue effects should be avoided by not offering more than 10 or so choice comparisons within a single experiment. This number, of course, may vary with the difficulty of the choices offered and the total length of the survey. If one can randomise the order in which pairs of alternatives are presented, the relative magnitudes of the model coefficients will not be biased. Such randomisation is greatly facilitated by computer-based interviewing.

For ranked data, the results suggest that one should not go beyond using the first three or four ranks as choices in exploded logit, that one should check the extent to which the results change as more ranks are used, and that scale factors should be estimated to avoid biased estimates. If possible, different blocks of alternatives should be given to different groups of respondents for ranking, and the ranking task should be administered in way which encourages respondents to give equal attention to the ranking of the more-preferred and less-preferred alternatives.

The types of experimental designs and interview methods used to collect the data in the two case studies presented here are typical of those used in many recent transport SP studies. While we cannot prove the general existence of rank order and fatigue effects based on these case studies alone, we can assert that all SP studies are at least *susceptible* to these effects. One should therefore (a) use blocked experimental designs and randomised task ordering to the greatest extent possible, (b) keep SP choice or ranking tasks as brief and stimulating as possible, and (c) avoid using exploded logit analysis with rank order data unless one explicitly estimates scale differences across the rankings.

Appendix A: Experimental designs for Case Study 1.

Experiment 1 – Bus service levels

Alternative	Fare	Punctuality	Interchanges	Travel time
Best	down by 20%	better	none	20% less
Worst	up by 20%	worse	one	current
A	current	better	one	20% less
B	current	current	none	20% less
C	current	worse	one	current
D	down by 20%	current	one	current
E	down by 20%	worse	none	20% less
F	down by 20%	better	one	20% less
G	up by 20%	worse	one	20% less
H	up by 20%	better	none	current
I	up by 20%	current	one	20% less

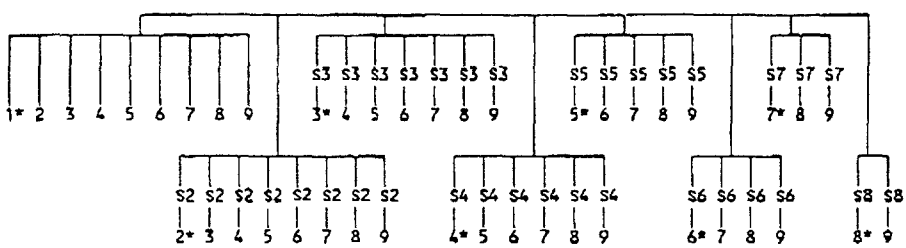
Experiment 2 – Bus stop facilities

Alternative	Fare	Ticketing	Information	Waiting
Best	current	automated	real time	shelter
Worst	up by 20%	current	current	none
A	current	current	real time	shelter
B	current	automated	current	none
C	current	automated	current	shelter
D	up by 20%	current	current	shelter
E	up by 20%	automated	current	shelter
F	up by 20%	automated	real time	none
G	up by 10%	current	current	none
H	up by 10%	automated	real time	shelter
I	up by 10%	automated	current	shelter

Experiment 3 – Bus vehicle factors

Alternative	Fare	Seat avail.	Cleanliness	Information
Best	current	available	clean in + out	destination sign
Worst	up by 40%	no seat 10 mn	current	none
A	current	available	current	destination sign
B	current	no seat 2 min	clean inside	none
C	current	no seat 10 min	clean in + out	destination sign
D	up by 40%	available	clean in + out	none
E	up by 40%	no seat 2 min	current	destination sign
F	up by 40%	no seat 10 min	clean inside	destination sign
G	up by 20%	available	clean inside	destination sign
H	up by 20%	no seat 2 min	clean in + out	destination sign
I	up by 20%	no seat 10 min	current	none

Appendix B: Artificial tree structure for Case Study 1.



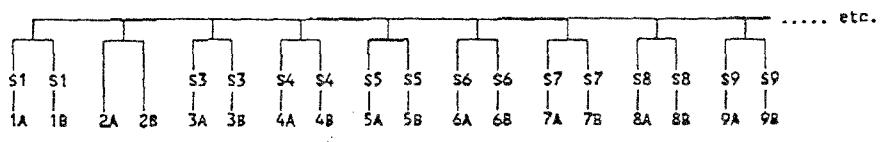
1–9: alternatives ranked 1 to 9
S2–S8: scale factors for ranks 2 to 8 (1 is base)
*: chosen alternative in "branch" of tree

Appendix C: Experimental levels for Case Study 2.

Variable		Actual car users	Actual train users
(A)	Rail fare	1 – Current level†	1 – Current level†
		2 – 10, 15, 20 or 25% lower*	2 – 10, 15, 20 or 25% higher*
		3 – 20, 30, 40 or 50% lower*	3 – 20, 30, 40 or 50% higher*
(B)	Journey time	1 – Current level†	1 – Current level†
		2 – 15% shorter	2 – 15% longer
		3 – 30% shorter	3 – 30% longer
(C)	Inter-changes	1 – Current number†	1 – Current number†
		2 – 1 less, if possible	2 – 1 more
		3 – 2 less, if possible	
(D)	Comfort level**	1 – Current level	1 – Current level
		2 – Improved level	2 – Worsened level

† = current level reported by respondent
* = set of percentages selected at random for each respondent
** = detailed verbal descriptions were used for comfort levels

Appendix D: Artificial tree structure for Case Study 2



1A–9A: alternative A for choices in sequence 1 to 9
1B–9B: alternative B for choices in sequence 1 to 9
S1–S9: scale factors for choices in sequence 1 to 9 (2 is base)

References

Bates J (1988) Econometric issues in SP analysis. *Journal of Transport Economics and Policy* V 12(1): 59–70.

Ben-Akiva M & Lerman S (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand.* MIT Press. Cambridge, Mass.

Ben-Akiva M & Morikawa T (1990) Estimation of travel demand models from multiple data sources. *11th International Symposium on Transportation and Traffic Theory.* Yokohama.

Ben-Akiva M, Morikawa T & Shiroishi F (1991) Analysis of the reliability of preference ranking data. *Journal of Business Research* V(23): 253–286.

Bonsall P (1985) Transfer price data – Its definition, collection and use. In: E Ampt, A Richardson & W Brog (eds), *New Survey Methods in Transport.* VNU Science Press. Utrecht.

Bradley M & Daly A (1992) Estimation of logit choice models using mixed stated preference and revealed preference information. *6th International Conference on Travel Behaviour.* Quebec. (Revised for forthcoming publication.)

Bradley MA, Grosvenor T & Bouma A (1988) An application of computer-based stated preference to study mode switching in the Netherlands. *PTRC 16th Annual Summer Meeting – Proceedings of Seminar D.* University of Bath.

Bradley MA & Kroes EP (1992) Forecasting issues in stated preference survey research. In E Ampt, A Richardson & A Meyburg (eds), *Selected Readings in Transport Survey Methodology.* Eucalyptus Press. Melbourne.

Daly AJ (1987) Estimation of 'Tree' Logit Models. *Transportation Research* 21B: 251–267.

Daly AJ (1990) Integration of revealed preference and stated preference date in model estimation. *Banff Invitational Symposium on Consumer Decision Making and Choice Behavior.* Alberta.

Daly AJ & de D Ortúzar J (1990) Forecasting and data aggregation: Theory and practice. *Traffic Engineering and Control* Dec. 1990: 632–643.

Hausman JA & Ruud PA (1987) Specifying and testing econometric models for rank-Ordered data. *Journal of Econometrics* 34: 83–104.

Hensher DA & Bradley M (1993) Using stated choice data to enrich revealed preference discrete choice models. *Marketing Letters* 4(2): 139–152.

Morikawa T, Ben-Akiva ME & Yamada K (1992) Estimation of mode choice models with serially correlated RP and SP data. *7th World Conference on Transport Research.* Lyon.

Morikawa T (1989) Incorporating stated preference data in travel demand analysis. Doctoral dissertation, MIT. Cambridge, Mass.

Morikawa T, Ben-Akiva M & McFadden D (1990) Incorporating psychometric data in econometric demand models. *Banff Invitational Symposium on Consumer Decision Making and Choice Behavior.* Alberta.

Swait J & Louviere JJ (1993) The role of the scale parameter in the estimation and comparison of multinomial logit models. *Journal of Marketing Research* (forthcoming).

Widlert S, Bradley M, Kroes E, Sheldon R, Garling T & Uhlin S (1989) Preferences for bus and underground services in Stockholm. *6th World Conference on Transportation Research.* Yokohama.