Title: Key Issues and Potential Solutions for Understanding Health Care Preference Heterogeneity Free from Patient Level Scale Confounds

Running title: Key Issues and Potential Solutions for Understanding Health Care Preference Heterogeneity

Authors:

Catharina G.M. Groothuis-Oudshoorn, PhD, University of Twente, Department of Health Technology and Services Research, PO Box 217, 7500 AE Enschede, The Netherlands

Terry N. Flynn, PhD, TF Choices LTD, Nottingham NG5 8JE, United Kingdom

Hong Il Yoo, PhD, Durham University, Durham University Business School, Durham DH1 3LB, United Kingdom

Jay Magidson, PhD, Statistical Innovations Inc., 375 Concord Ave. (Suite 007), Belmont Massachusetts, 02478

Mark Oppe, PhD, EuroQol Research Foundation, Marten Meesweg 107, 3068 AV Rotterdam, The Netherlands

Corresponding author:

Catharina G.M. Groothuis-Oudshoorn, PhD, University of Twente, Department of Health Technology and Services Research, Faculty BMS, PO Box 217, 7500 AE Enschede, The Netherlands, Phone: 0031 53 489 5374, E-mail: c.g.m.oudshoorn@utwente.nl

1. Introduction

Health care is becoming increasingly personalized to individuals. For instance, the United States Food and Drug Administration (FDA) attempts to incorporate patient preferences into regulatory decision-making and is willing to approve treatments even if the benefit-risk profile is acceptable only to a segment of risk-tolerant patients[1]. Countries with extra-welfarist health care systems (relying on population preferences) now recognize that heterogeneity in individual preferences may be important conceptually in addressing issues such as child health, social care-related quality of life and carer well-being.

Identifying and understanding heterogeneity in preferences can be helpful for identifying patients' needs, tailoring treatments to certain patient subgroups, and developing decision support tools in health care[2]. Therefore, studying preference heterogeneity among patients is meaningful. Segmenting patient and physician populations can lead to better understanding of the diversity of needs and desires among segments, and to providing better treatments, prevention programs, health services, and products [3][4].

Whilst preference heterogeneity in continuous outcome variable models (like the Time Trade-Off – TTO) has been investigated for many years, issues remain. Furthermore, increasing use of limited dependent variable (LDV, such as logit and probit) models raise new, crucial issues. This paper provides a summary of discussion in a half-day symposium attached to the IAHPR 2017 Glasgow conference dealing with such issues. The aim is to raise awareness of some crucial under-appreciated issues in both TTO and LDV models: it will not attempt to "police" the fields but merely introduce the key issues and discuss advantages and disadvantages of potential solutions.

2. Definition of heterogeneity and how traditional solutions can fail

McFadden's [5] random utility model allows the researcher to conceptualize two major types of interpersonal preference heterogeneity: heterogeneity in structural preference parameters (real trade-offs) and heterogeneity in the overall (variance) scale of utility. Classical modeling approaches focus on heterogeneity in the structural parameters, or the idea that different people make different tradeoffs across product attributes. Early studies captured observed heterogeneity by allowing the structural parameters to vary with the decision maker's observed characteristics, such as health and income. But preference heterogeneity may be expected to exist even among people with an identical sociodemographic profile, and capturing unobserved heterogeneity has become a central theme of the modern literature. A seminal study by Kamakura and Russell [6] introduces the latent class logit model (LCL) which postulates that there are C distinct types of decision makers for some finite number C, where each type or preference segment has its own structural parameters; this is a model of unobserved heterogeneity since each person's segment membership is probabilistic. Revelt and Train [7] have popularized an alternative method known as the mixed logit model (MXL); the inspiring

metaphor is that each person has their own structural parameters which are like realizations of continuous random variables, and the population preference distribution is akin to a joint distribution of such variables. McFadden and Train [8] developed a unifying framework that accommodates both LCL and MXL, by casting LCL as a cousin of MXL which specifies the structural parameters as discrete random variables instead of continuous random variables.

Heterogeneity in the scale parameter (i.e. heteroscedasticity) is a reduced form representing multiple latent processes, such as heterogeneous preferences for unobserved product attributes and heterogeneity in the decision maker's attentiveness during choice experiments [9]. A linear regression model, combined with suitably robust standard errors, allows the researcher to be agnostic about scale heterogeneity when studying the structural parameters. In a non-linear LDV model, however, neglected scale heterogeneity induces misspecification bias, meaning that one may draw false conclusions regarding the structural parameters unless scale heterogeneity is explicitly specified and estimated – in other words, regression estimates may be *biased* in an unknown direction and with unknown magnitude. This issue was first identified by Yatchew and Grilisces [10], and empirically illustrated in the subsequent literature. For example, Hensher and Louviere [11] found that most of observed heterogeneity in the structural parameters vanishes once models account for observed scale heterogeneity. Fiebig et al. [9] found that accounting for unobserved scale heterogeneity affects posterior inferences on unobserved heterogeneity in the structural parameters substantively. Although scale heterogeneity is a prevalent issue in healthcare DCEs, only a few published studies pay attention to it and even fewer use formal methods to identify and account for the impact of it [12] .

3. Discussion of more complex heterogeneity in various models

State-of-the-art models that account for both sources of interpersonal heterogeneity (heterogeneous structural parameters and scale heterogeneity) attempt to segment/cluster respondents and therefore deal with heterogeneity in preferences were presented at the symposium. As an introduction, the first presentation clarified the motivation and intuition behind this new generation of modeling approaches; two patterns of choices, although looking quite different in terms of the sensitivity to stimuli, might be driven by the same underlying (true, mean) preferences. Their different choice consistency (variances) might merely reflect differing engagement with the task rooted in education levels, age, etcetera which contribute to scale heterogeneity. To minimize possible confounding, the researcher must specify and estimate both sources of interpersonal heterogeneity simultaneously.

The second presentation focused on the scale-adjusted latent class analysis (SALC) model, a modern extension of LCL that was originally proposed by Magidson and Vermunt [13]. Under SALC, the population not only comprise C preference segments but also K scale segments for some finite number K, where each scale segment has its own scale parameter much as each preference segment has its own structural parameters. An individual decision maker is simultaneously a member of one

preference segment and one scale segment, and their membership in either segment is probabilistic. Importantly, a person's membership in a particular preference segment does not preclude their membership in any scale segment. The SALC model thus allows people who have identical structural parameters to have non-identical scale parameters, thereby exhibiting seemingly different choice patterns in raw data.

The third presentation focused on the generalized multinomial logit model (GMNL) of Fiebig et al. [9], a modern extension of MXL. In simple, GMNL is to MXL what SALC is to LCL. Under GMNL, an individual's preferences are modeled as draws from a continuous distribution of random structural parameters as well as that of random scale parameters. The model thus offers qualitatively the same kind of flexibility as SALC, since people who (metaphorically) happen to draw identical structural parameters may still draw different scale parameters. Indeed, within the unifying framework of McFadden and Train [8], the difference between GMNL and SALC comes down to the use of continuous vs discrete distributions to capture population heterogeneity.

Finally, the last presentation focused on cluster analysis of TTO data with a hierarchical clustering algorithm [14]. The TTO research was interesting in that it might be described as taking elements of both cluster-based and continuous distributions (although using a continuous outcome). It showed that different segments of the population use different parts of the TTO scale, and moreover, distribute similar health states differently across the TTO scale: more or less evenly spaced on the TTO scale versus concentrated at the top and the bottom of the TTO scale. This is indicating that the functional form underlying their choice function may differ. Separation of these groups becomes important in understanding better how different groups might react to a TTO task, and how simple aggregation might not always be appropriate.

4. Discussion of advantages and disadvantages of various models

It is incumbent upon the analyst to decide between competing models and this formed one topic for discussion. It was re-iterated that the mean-confound is perfect and there is simply no way for the analyst to know with certainty if the correct split has been performed. The SALC's strengths are its ability to seek out a set of "more likely" splits. Although well-known statistical criteria (such as a BIC) are frequently used to decide on the optimal model, discussion arose on how all such models tend to overfit, giving more segments than theory and common sense would suggest. Knowledge of the types of citizen typically present in large population studies was shown to help in choosing a final model that was quite parsimonious.

A related issue pertinent to LCL and SALC is how to determine the number of relevant classes. The standard information criteria tend to overstate the number, to include minor classes that are not policy relevant. A major new development in LCL is the ability to utilize a hierarchical tree to structure the latent classes with major policy-relevant classes being identified at the root of the tree and then split

further as needed to explain all the heterogeneity [15]. This innovative methodology was extended to SALC models by one of the presenters.

An issue common to GMNL/MIXL and SALC models is that the stated maximum of the likelihood function may not, in fact, be the global maximum. Various solutions were discussed, depending on the analysis type but there was general agreement that simply using multiple sets of starting values might not be sufficient. The GMNL/MIXL model solution was innovative in using global search algorithms that were less prone to finding inferior local maxima [16], whilst alternative variables as covariates (constructed in an entirely different manner) from BWS data were proposed for SALC models with one author finding that these measures, published in detail elsewhere, are very powerful in avoiding local maxima [17].

Another discussion issue concerned the design of the study. There was clear agreement that a bad design can't be solved by any of these techniques. There was less agreement that for good designs "methods don't matter": some participants expressed a view that *proper* interrogation of the data by any of these methods can give you broadly similar results, whilst others expressed concerns at multi-variate normal distributions assumed in MXL/GMNL models with a feeling that "fewer assumptions is better", particular among supporters of the TTO and SALC. However, and most crucially, the discussion seemed to converge upon the need to "know one's data" far better than has hitherto been the case. For instance BWS is unequivocally an individual-level model and "looking at the data first" is a prerequisite. In various instances, variance effects might not be to do with a respondent's cognitive abilities but might, in fact, reflect incorrect information. An example was given of high consistency but upon outcome combinations that make no sense, given the application. The general point seemed to be that if the analyst really understands how respondents were reacting to the experiment, and interacting with it, then analysis decisions flow fairly naturally.

The need to understand the data flowed in various directions, with calls for better qualitative research to observe the choice process and the need for more than one data source. Quantitative analysts might be more comfortable with simply obtaining more experimental econometric data – collected from a second preference study, since it has been known since at least 1993 that two sets of data in principle allows separate estimation of the two unknowns (the mean and variance vectors – assuming only two scale parameters since the number of unknowns must not exceed the number of data sources)[18]. The use of attitudes as a second source was discussed; a potential advantage is that these, if collected appropriately, might be considered to be more robust (to the econometrician anyway) and less subjective. Thus, an *attitude* such as "general views on conventional doctor-prescribed medicines" might provide valuable insights into a *preference* for a particular medicine over another (type of) intervention.

The flexibility of the SALC model, with the current version 5.1 of the Latent GOLD software [19] being user-friendly (being largely GUI based) in dealing with the scale parameter was demonstrated and discussed. This reflected recognition that choice consistency (variances) may differ between ranks (for instance between a best and worst choice) and/or between classes, giving great flexibility overall as to heteroscedasticity on the latent scale. The discussion touched upon the issue that if using the psychology paradigm which assumes people are not perfectly deterministic (consistent) on repeated occasions then this flexibility is important. For example BWS is intrinsically an individual level model.

On the other hand, some participants were more comfortable in the welfarist economics conceptualization of random utility theory (assuming that errors are simply factors the analyst couldn't observe) and downplaying the role of "people making errors". Whilst such fundamental differences were never going to be resolved, it was agreed that far better explorative and confirmatory studies were required (where possible).

5. Conclusion

The discussion summarized here was stimulating and achieved its aims of setting out relative advantages and disadvantages of both segmentation and continuous probability distribution approaches in LDV approaches, together with new issues in the TTO. It was particularly useful in making participants fully aware that the mean-variance confound in all ranking/DCE/BWS studies is perfect: none of the approaches used in practice can unequivocally claim correct separation of the two.

What was most useful was discussion of the tools the analyst might use to help minimize the chance of producing a grossly incorrect and/or misleading separation. Although certain issues such as choices of starting values for a maximization algorithm have the potential to bewilder the average applied practitioner, intuitive explanations abounded, including easy statistical methods and checks that although simple, are not typically "in the analyst's toolbox" were described to help identify spurious solutions from scale adjusted models. However, and reassuringly, strong messages from the session pertained to the need for common sense and theory, reinforcing a general desire expressed for practitioners to look at their data, no matter what analysis method they ultimately used. It is hoped that such sessions become more common – at the very least this one provided a much needed appreciation of competing methods' advantages and disadvantages which should help health analysts considerably.

Compliance with Ethical Standards

Funding: No funding was obtained for the writing of this commentary.

Conflicts of interest: Groothuis-Oudshoorn, Yoo and Oppe indicated that they have no potential conflicts of interest, relevant to the present study. Magidson declared having a financial interest in the Latent Gold software (as owner of Statistical Innovations). Flynn is owner of TF Choices which relies on the use of latent gold in commercial studies

Ethical approval: Not applicable

Informed consent: Not applicable

6. References

1. Johnson FR, Beusterien K, Özdemir S, Wilson L. Giving Patients a Meaningful Voice in United States Regulatory Decision Making: The Role for Health Preference Research. Patient - Patient-Centered Outcomes Res [Internet]. 2017;10:523–6. Available from: https://doi.org/10.1007/s40271-017-0250-z

2. Whitty JA, Fraenkel L, Saigal CS, Groothuis-Oudshoorn CGM, Regier DA, Marshall DA. Assessment of Individual Patient Preferences to Inform Clinical Practice. Patient - Patient-Centered Outcomes Res [Internet]. 2017;10:519–21. Available from: https://doi.org/10.1007/s40271-017-0254-8

3. Craig BM, Lancsar E, Mühlbacher AC, Brown DS, Ostermann J. Health Preference Research: An Overview. Patient - Patient-Centered Outcomes Res [Internet]. 2017;10:507–10. Available from: https://doi.org/10.1007/s40271-017-0253-9

4. Deal K. Segmenting Patients and Physicians Using Preferences from Discrete Choice Experiments. Patient - Patient-Centered Outcomes Res [Internet]. 2014;7:5–21. Available from: https://doi.org/10.1007/s40271-013-0037-9

5. McFadden D. Conditional logit analysis of qualitative choice behavior. In: Zarembka P, editor. Front Econom. New York: Academic Press; 1974. p. 105–42.

6. Kamakura WA, Russell G. A probabilistic choice model for market segmentation and elasticity structure. J Mark Res. 1989;26:379–90.

7. Revelt D, Train K. Mixed Logit With Repeated Choices: Households' Choices Of Appliance Efficiency Level. Rev Econ Stat [Internet]. 1998;80:647–57. Available from: https://econpapers.repec.org/RePEc:tpr:restat:v:80:y:1998:i:4:p:647-657

8. McFadden D, Train K. Mixed MNL models of discrete response. J Appl Econom. 2000;15:447–70.

9. Fiebig D, Keane M, Louviere J, Wasim N. The generalized multinomial logit model: accounting for scale and coefficient heterogeneity. Mark Sci. 2010;29:393–421.

10. Yatchew A, Griliches Z. Specification error in probit models. Rev Econ Stat. 1985;67:134–9.

11. Hensher DA, Louviere J. Combining sources of preference data. J Econom. 1999;89:197–221.

12. Wright SJ, Vass CM, Sim G, Burton M, Fiebig DG, Payne K. Accounting for scale heterogeneity in healthcare-related discrete choice experiments when comparing stated preferences: a systematic review. Patient - Patient-Centered Outcomes Res. 2018;

13. Magidson J, Vermunt JK. Removing the Scale Factor Confound in Multinomial Logit Choice

Models to Obtain Better Estimates of Preference. Sawtooth Softw Conf Proc. 2007. p. 139–54.

14. Mardia KV, Kent JT, Bibby JM. Cluster analysis. Multivar Anal. London: Academic Press; 1979. p. 360–93.

15. Van den Bergh M, van Kollenburg GH, Vermunt JK. Deciding on the starting number of classes of a latent class tree. Sociol Methodol. 2018;in press.

16. Hole AR, Yoo HI. The Use of Heuristic Optimization Algorithms to Facilitate Maximum Simulated Likelihood Estimation of Random Parameter Logit Models. J R Stat Sociiety Ser C. 2017;66:997–1013.

17. Louviere J, Flynn TN, Marley AAJ. Best-worst scaling: theory, methods and applications. 1st ed. Cambridge: Cambridge University Press; 2015.

18. Swait J, Louviere J. The Role of the Scale Parameter in the Estimation and Comparison of Multinomial Logit Models. J Mark Res [Internet]. American Marketing Association; 1993;30:305–14. Available from: http://www.jstor.org/stable/3172883

19. Vermunt JK, Magidson J. Technical Guide for Latent Gold 5.1: Basic, Advanced, and Syntax. Belmont, MA: Statistical Innovations Inc.; 2016.