VALUE IN HEALTH 22 (2019) 157–160

Available online at www.sciencedirect.com

ScienceDirect

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

Brief Report

The Internal Validity of Discrete Choice Experiment Data: A Testing Tool for Quantitative Assessments

F. Reed Johnson, PhD[*], *Jui-Chen Yang, MEM, Shelby D. Reed, PhD, RPh*

Preference Evaluation Research (PrefER) Group, Duke Clinical Research Institute, Duke University, Durham, NC, USA

ABSTRACT

Objectives: To develop a tool for testing internal validity of discrete choice experiment (DCE) data, deploy the program, and collect summary test results from a sample of active health researchers to demonstrate the practical utility of the tool in a wide range of health applications. **Methods:** A previously developed Gauss program had been in use for testing internal validity. The program was translated to MATLAB and adapted, compiled, and deployed. Sixty-seven authors who had coauthored one or more published DCE studies between 2013 and 2016 were contacted by email; provided access to the tool, instructions, and an example data file; and invited to submit test summaries for tabulation. **Results:** Twenty-one researchers from 10 countries contributed test results from a total of 55 DCE data sets. Fifty-one studies included at least two out of a possible six tests. Attribute dominance was the most common test, and stability had the highest failure incidence. Only three summaries included a transitivity test, and no failures were detected. **Conclusions:** It was possible to evaluate multiple internal validity checks for most data sets even when the experimental design did not explicitly include tests. Nevertheless, internal validity is rarely reported. Free availability of the tool for testing data quality could improve both reporting and more careful design of DCE studies to help validate and interpret stated preference data.

Keywords: discrete choice experiments, health preferences, testing tool, validity testing

Introduction

Although there are indications of a growing consensus on good practice in several areas of health preference research [1–3], there are no widely accepted indicators of data quality, such as internal validity. A recent call for research to advance the science of patient preference assessment in regulatory submission and review in the United States highlights the lack of consensus on the meaning of validity and the need for tools to better assess the quality of preference data [4]. In an effort to fill this gap, our objectives were 1) to identify a set of internal validity tests for discrete choice experiment (DCE) data in health preference research, including logic, consistency, and conformity with the conceptual framework; 2) to develop and make available a free internal validity test tool that preference researchers and regulators can use to assist in evaluating data quality in past and future studies; and 3) to collect, compile, summarize, and disseminate the results of a database of internal validity tests for DCE applications in health and health care.

Internal-validity tests of DCE data are checks on the logic, consistency, and trade-off assumptions of recorded responses to DCE questions [5]. Because DCE respondents complete a series of related choice questions, the data present an opportunity to evaluate respondent-specific response patterns. There are several potential internal validity tests, depending on the structure of the DCE design. Some tests must be explicitly embedded in an experimental design, whereas others can occur in any design. We describe six different internal validity tests here.

Methods

Internal Validity Tests

Stability (repeated question)

Preference researchers may repeat a choice question later in the sequence to check whether the respondent chooses the same alternative.

* *Address correspondence to:* F. Reed Johnson, Departments of Population Health Science and Medicine, Preference Evaluation Research (PrefER) Group, Duke Clinical Research Institute, Duke University, P.O. Box 17969, Durham, NC 27715.

E-mail: reed.johnson@duke.edu

1098-3015/$36.00 - see front matter

https://doi.org/10.1016/j.jval.2018.07.876

Within-set dominated pairs
Preference researchers may include a choice question when one alternative is unambiguously better for all attributes to check whether the respondent chooses the dominated alternative within the set.

Across-set dominated pairs
The across-set dominated-pairs test is a generalization of the within-set dominated-pairs test. Although across-set dominated-pairs could occur naturally in an experimental design, many such comparisons can occur for designs with a constant status quo or opt-out alternative with fixed attribute levels. Suppose there are two choice sets, 1 and 2. A_1 and B_1 designate attribute-level profiles for alternatives A and B in set 1. A_2 and B_2 designate attribute-level profiles in set 2. The alternative C profile is common to both choice sets.

Choice set 1: $A_1 \sim B_1 \sim C$
Choice set 2: $A_2 \sim B_2 \sim C$

Suppose alternative B is chosen in choice set 1, indicating that B_1 is preferred to both A_1 and C. If the attributes in B are naturally ordered and B_1 is inferior to B_2 for every attribute, then choosing C in choice set 2 fails a logic test; if B_1 is preferred to C in choice set 1, then B_2 must be even more attractive than C in choice set 2. Alternatively, if C is chosen in choice set 2, then logic prohibits choosing B_1 in choice set 1 because C was shown to be preferred to the better levels of B_2.

Transitivity
Transitivity is a requirement for rational preferences and is one of the fundamental axioms of economic utility theory. Transitivity tests will not occur naturally in efficient experimental designs. Preference researchers must design a transitivity test separately from the experimental design question sequence. Transitivity requires that if alternative X is preferred to alternative Y and alternative Y is preferred to alternative Z, then alternative X must be preferred to alternative Z.

Attribute dominance (noncompensatory preferences)
Choice experiments assume that respondents have compensatory preferences; they should be willing to accept a reduction in one desirable attribute in return for a sufficiently large compensating increase in another desirable attribute. Attribute dominance is an observed noncompensatory pattern in which respondents choose the alternative with the better level of one attribute in all or nearly all choice questions.

Straight-lining or flat-lining
The probabilities that the more preferred alternative will always appear in the same position for 5, 6, or 7 or in more randomly positioned pairwise comparisons are 3%, 1.5%, and less than 1%, respectively. The likelihoods are small enough to suggest that respondents who always or nearly always choose the same alternative in the same position are not evaluating the alternatives carefully. In survey research, such patterned responses are called straight-lining or flat-lining.

MATLAB Internal Validity Test Tool

To facilitate applying, reporting, and evaluating internal validity tests in preference studies, we developed a freely available MATLAB internal validity test tool. Six preference researchers assisted with beta-testing early versions of the MATLAB tool in March and April 2017. The tool performs the internal-validity tests described earlier. The tool can be run directly on a Windows computer with MATLAB installed, or it can be run on a Windows computer after installing the free MATLAB Runtime module. Links to the program, instruction file, an example data file, as well as an example test summary are provided in the Supplemental Materials found at https://doi.org/10.1016/j.jval.2018.07.876.

Data Collection

Data collection was initiated by an email on April 24, 2017, from Martin Ho, Associate Director for Quantitative Innovation, US Food and Drug Administration's Center for Devices and Radiological Health, indicating Food and Drug Administration's support for the study. The email was sent to a list of 67 researchers who had published health preference studies in the previous 4 years. Between May and October 2017 we sent several follow-up emails that contained the links to the MATLAB tool, related documentation, and a background survey to encourage researchers to participate in the study. Results of the background survey are not reported in this article.

The emails requested that researchers run the tool on dce data from their previous preference studies and return the output files to us. to increase participation, researchers also had the option of just sending their dce data to us. in those cases, we ran the tool on the anonymized data sets, sent the results to the researcher, and deleted the dce data. in mid-may 2017, we also offered a free webinar to demonstrate use of the tool, to assist in interpreting the test results, and to answer any questions relating to the tool. to ensure anonymity, each study was assigned one or more test-summary ids, depending on whether more than one set of test results was related to the same study. Only the summaries of the test results were retained for analysis. We report results for all summaries submitted by the end of december 2017.

Results

In total, 21 researchers from 18 research groups in Argentina, Australia, Canada, Denmark, Germany, the Netherlands, the United Kingdom, the United States, and Vietnam contributed test results from a total of 55 DCE data sets. Seven researchers provided summaries for more than one data set from the same study. The studies covered a wide range of health applications.

Table 1 presents descriptive statistics for the test summaries. After accounting for straight-lining, which can be evaluated in any dce, 51 of 55 studies included at least two tests. Table 2 presents

Table 1 – Summary of test summaries received (N = 55)

Study feature	Range	Mean ± SD	Median
Number of respondents	14–3003	549 ± 511	486
Number of questions	5–24	11.8 ± 4.8	10
Number of alternatives	2–4	2.4 ± 0.5	2
Number of attributes	3–9	5.3 ± 1.3	5
Minimum number of attribute levels	2–5	2.9 ± 1.0	3
Maximum number of attribute levels	2–20	4.3 ± 2.4	4

Table 2 – Internal validity test results (N = 55)

Test Type	Number of summaries with the test	% Failures				
		Mean ± SD	Median	25th Percentile	75th Percentile	Range
Stability (repeated question)	16	30% ± 26%	24%	10%	44%	0%–81%
Within-set dominated pairs	21	18% ± 20%	7%	3%	27%	0.2%–58%
Across-set dominated pairs	28	6% ± 9%	2%	1%	7%	0%–29%
Transitivity	3	0% ± NA	0%	0%	0%	0%–0%
Attribute dominance	42	22% ± 14%	20%	11%	35%	0.3%–50%
Straight-lining or flat-lining	55	7% ± 11%	2%	1%	8%	0.2%–43%

NA, not applicable.

the results from the 55 test summaries. attribute dominance was the most common test, and stability had the highest failure incidence. Only three summaries included a transitivity test and no failures were detected.

Discussion

Nearly all the submitted test summaries (51 of 55) contained at least two of six possible internal-validity tests. Nevertheless, results of such tests are rarely reported [6]. Without routine reporting of results from internal-validity tests, it is difficult for researchers to gauge the extent to which their DCE data quality compares with other published studies.

It is important to understand that the conceptual framework of preference research is based on random-utility theory, which postulates that utility is latent and cannot be observed without error [7]. Its randomness may be caused by the following:

1. *Excluded attributes:* The complete list of attributes that affect utility may not be known or experimentally controlled.
2. *Unobserved taste variations:* Unobservable respondent characteristics may influence utility.
3. *Measurement error:* Many factors can introduce measurement error into DCE data. The most important of these relate to the construction of the survey instrument or the mode of administration, including the following:
   - ambiguities in and reading level of attribute definitions;
   - complexity of choice questions;
   - number of choice questions;
   - features of the experimental design; and
   - fatigue or inattention of respondents.

Random utility does not assert that people's underlying preferences are random, only that researchers are unable to observe utility accurately. Apparent internal validity test failures could be the result of behavioral considerations such as learning, fatigue, or simplified heuristics to minimize cognitive effort that could indicate problems related to the design of the survey instrument.

Given that random selection of alternatives could pass (or fail) the tests about half the time, stability tests or within-set dominated pairs are relatively weak tests of internal validity. The reliability of the repeated-question test can be influenced by how proximate the two questions are in the sequence. Also, when choice questions have alternatives with very similar utility perceived by respondents, either alternative could be quite acceptable. Nevertheless, choosing one over the other(s) could be classified as a validity failure in a repeated-question or other test.

Attribute dominance may be explained by the following competing hypotheses. One is that some respondents value a given attribute very highly and the range of variation in other attributes does not offer sufficient compensation for trading away from the next-to-worst level of that attribute. Another is a simplified decision heuristic that allows the respondent to avoid expending effort to evaluate the trade-offs shown. In cases in which researchers constrain attribute levels to gain greater precision around relatively narrow, clinically relevant ranges, the prevalence of respondents dominating on a given attribute is likely to be greater than in cases in which wider ranges are chosen to induce trade offs among attributes. Conversely, dominance failures can be identified only for naturally ordered attributes and thus could undercount failures if they occur in categorical attributes having no natural order but still are ranked by respondents.

Apparent straight-lining or flat-lining could occur in a DCE with a fixed alternative always shown in the same position. In this case, respondents could be indicating a reasonable preference for the constant opt-out or status quo alternative. The test summary tabulates the number of times each alternative was chosen, so this pattern would be evident if it occurred.

Thus, preference researchers should avoid the temptation to simply count validity results in a data set to derive an index of quality and avoid simplistic reporting of internal validity test failures without further examination. Although we have characterized responses that do not conform to typical expectations about consistency, logic, and utility theoretic assumptions as failures, there well may be plausible explanations for such responses, and researchers should take care in considering how to evaluate and interpret internal validity test results. Failing any of the tests does not necessarily mean data are uninformative about preferences, and findings from validity testing must be interpreted with regard to the features and design of each DCE. Researchers should determine the extent to which results are sensitive to inclusion or exclusion of suspect observations; the demographic, educational attainment, and disease experience characteristics of respondents who fail one or more tests; and any failure pattern that suggests problems with the instrument itself.

Furthermore, scrutiny of apparent internal validity test failures can yield insights into the quality of the instrument used to elicit preferences as well as insights into respondent preferences other than those obtained from mechanical statistical analysis of the data. For example, we recently found a surprisingly high rate of failure on a dominated-pair test in one of our patient preference studies for a progressive, severely debilitating disease. On closer examination of the data, we found evidence of a "value of hope" or value of "doing something" associated with choosing an active treatment alternative rather than a no treatment alternative. In further analysis, we were able to estimate that value separately from the value associated with the attributes in the DCE design [8].

Although prospective inclusion of internal validity testing represents good practice in health preference studies, researchers should recognize that such tests impose a penalty on statistical

precision. For individuals with stable, logical preferences, researchers do not gain additional statistical information from repeated questions, dominated pairs, or transitivity tests. Tests of transitivity requiring comparisons of XY, YZ, and XZ are not generated in any D-efficient experimental design because no additional statistical information would be gained from such questions. Including alternatives in the experimental design in which alternatives share a common attribute level (attribute overlap) encourages respondents to evaluate other attributes when making a choice. Although this reduces observed dominance, it also reduces the statistical efficiency of the design. It also is not clear what threshold should be used to classify dominance failure. The default used in the tool is 75%, but users can specify any value. We encourage researchers to analyze their data using the tool's full tabulation of the number of choices for the better level of all ordered attributes and to evaluate possible reasons for the observed pattern and sensitivity of their results when applying different thresholds to identify attribute dominance.

We must acknowledge that the data reported here have limited generalizability. First, we obtained only 55 self-selected data summaries from a much larger list of published studies. Nevertheless, the studies in the database are fairly similar when compared with reviews published by Marshall et al. [9] covering studies published between 2005 and 2008 and Clark et al. [10] covering studies published between 2001 and 2013. The generalizability of our results also is limited by the small number of submissions we received from some groups most active in health preference research. Furthermore, although we hoped that assurances of anonymity would encourage researchers to submit unfavorable test results, we cannot assume that submissions were unfiltered. We also do not know whether a test summary was based on a full initial study sample or on a cleaned final analysis sample. There also remains the possibility that researchers may have incorrectly specified information about their study when running the MATLAB tool, in which case some findings may have misidentified internal-validity failures. Finally, some studies contributed more than one test summary and some researchers contributed more than one study, both of which can be problematic when it comes to interpreting the results, because we treated each test summary as an independent observation. Controlling for correlation due to multiple test results from the same study and/or multiple studies from the same researcher would require statistical modeling, which is beyond the scope of this descriptive report.

Conclusions

The most significant contribution of this study was to widely deploy a free tool for testing the internal validity of DCE data. The tool was made available to most practicing preference researchers and can be used to assess DCE data quality in future studies. We recommend that study sponsors, regulators, and journal editors encourage researchers to use internal validity testing as part of assessments of data quality. We also are hopeful that the tool will facilitate including more internal-validity tests in DCE designs and reporting their test results and evaluations in journal publications and regulatory submissions.

Acknowledgments

We are grateful to the 21 researchers from Argentina, Australia, Canada, Denmark, Germany, The Netherlands, the United Kingdom, the United States, and Vietnam who generously contributed their studies to this report. We also are grateful to (*in alphabetical order*) Benjamin Craig, Angelyn Fairchild, Juan Marcos Gonzalez, Ellen Janssen, Marcel Jonker, Karen MacDonald, Deborah Marshall, Axel Mühlbacher, Jan Ostermann, Andrew Sadler, and Semra Özdemir for their patience in testing early, very buggy, versions of the MATLAB application.

Source of financial support: This project was supported by a subcontract to the Duke Clinical Research Institute on a grant from the US Food and Drug Administration to the Duke Margolis Center for Health Policy (grant no. U13FD005197).

Supplemental Materials

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.jval.2018.07.876.

REFERENCES

[1] Bridges JFP, Hauber AB, Marshall D, et al. Conjoint analysis applications in health—a checklist: a report of the ISPOR Good Research Practices for Conjoint Analysis Task Force. Value Health 2011;14:403–13.
[2] Johnson FR, Lancsar E, Marshall D, et al. Constructing experimental designs for discrete-choice experiments: report of the ISPOR Conjoint Analysis Experimental Design Good Research Practices Task Force. Value Health 2013;16:3–13.
[3] Hauber AB, Gonzalez JM, Groothuis-Oudshoorn CG, et al. Statistical methods for the analysis of discrete choice experiments: a report of the ISPOR Conjoint Analysis Good Research Practices Task Force. Value Health 2016;19:300–15.
[4] Levitan B, Hauber AB, Damiano MG, et al. The ball is in your court: agenda for research to advance the science of patient preferences in the regulatory review of medical devices in the United States. Patient 2017;10:531–6.
[5] Johnson FR, Bingham MF. Evaluating the validity of stated-preference estimates of health values. Swiss J Econ Stat 2001;137:49–63.
[6] Janssen EM, Marshall DA, Hauber AB, Bridges JFP. Improving the quality of discrete-choice experiments in health: How can we assess validity and reliability? Expert Rev Pharmacoecon Outcomes Res 2017;17:531–42.
[7] Manski CF. Structure of random utility models. Theory Decis 1977;8:229–54.
[8] Yang J-C, Johnson FR, DiSantostefano RL, Reed SD, Streffer J, Levitan B. Something is Better than Nothing: The Value of Active Intervention in Stated Preferences for Treatments to Delay Onset of Alzheimer's Disease Symptoms. Podium presentation at the International Society for Pharmacoeconomics and Outcomes Research, Baltimore MD, May 19-23, 2018. https://www.ispor.org/docs/default-source/presentations/1390.pdf?sfvrsn=a588211b_1 (accessed 9/20/2018).
[9] Marshall D, Bridges JF, Hauber B, et al. Conjoint analysis applications in health—How are studies being designed and reported? An update on current practice in the published literature between 2005 and 2008. Patient 2010;3:249–56.
[10] Clark MD, Determann D, Petrou S, et al. Discrete choice experiments in health economics: a review of the literature. Pharmacoeconomics 2014;32:883–902.