

mda

methods, data, analyses

JOURNAL FOR QUANTITATIVE METHODS AND SURVEY METHODOLOGY

Volume 16, 2022 | 2

Vignette Analysis: Methodology and Recent Developments

Lena M. Verneuer-Emre, Stefanie Eifler & Hermann Dülmer (Editors)

- | | |
|----------------------------------|---|
| Edgar Treischl & Tobias Wolbring | The Past, Present and Future of Factorial Survey Experiments |
| Clemens Maria Schmidt | Controlling for Taste Preferences |
| Stefanie Eifler & Knut Petzold | Fear of the Dark? |
| Sophie Cassel et al. | The Impact of Presentation Format on Conjoint Designs |
| Mengyao Hu et al. | Improving Anchoring Vignette Methodology in Health Surveys with Image Vignettes |
| Julia Kleinewiese | New Methodical Findings on D-Efficient Factorial Survey Designs |
| Alexander W. Schmidt-Catran | Factorial Surveys with Multiple Ratings per Vignette |

methods, data, analyses is published by GESIS – Leibniz Institute for the Social Sciences.

Editors: Melanie Revilla (Barcelona, editor-in-chief), Annelies Blom (Mannheim), Eldad Davidov (Cologne/Zurich), Edith de Leeuw (Utrecht), Gabriele Durrant (Southampton), Sabine Häder (Mannheim), Jan Karem Höhne (Duisburg-Essen), Peter Lugtig (Utrecht), Jochen Mayerl (Chemnitz), Norbert Schwarz (Los Angeles)

Advisory board: Andreas Diekmann (Leipzig), Udo Kelle (Hamburg), Bärbel Knäuper (Montreal), Dagmar Krebs (Giessen), Frauke Kreuter (Mannheim), Christof Wolf (Mannheim)

Managing editor: Sabine Häder
GESIS – Leibniz Institute for the Social Sciences
PO Box 12 21 55
68072 Mannheim
Germany
Tel.: + 49.621.1246526
E-mail: mda@gesis.org
Internet: <https://mda.gesis.org>

Methods, data, analyses (mda) publishes research on all questions important to quantitative methods, with a special emphasis on survey methodology. In spite of this focus we welcome contributions on other methodological aspects. We especially invite authors to submit articles extending the profession's knowledge on the science of surveys, be it on data collection, measurement, or data analysis and statistics. We also welcome applied papers that deal with the use of quantitative methods in practice, with teaching quantitative methods, or that present the use of a particular state-of-the-art method using an example for illustration.

All papers submitted to mda will first be screened by the editors for general suitability and then double-blindly reviewed by at least two reviewers. Mda appears in two regular issues per year (January and July).

Layout: Bettina Zacharias (GESIS)
Print: Bonifatius Druck GmbH Paderborn, Germany

ISSN 1864-6956 (Print)
ISSN 2190-4936 (Online)

© of the compilation GESIS, Mannheim, July 2022

All content is distributed under the terms of the Creative Commons Attribution 3.0 License. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

Content

- 133 Editorial:
Vignette Analysis: Methodology and Recent Developments
Lena M. Verneuer-Emre, Stefanie Eifler & Hermann Dülmer
-

RESEARCH REPORTS

- 141 The Past, Present and Future of Factorial Survey
Experiments: A Review for the Social Sciences
Edgar Treischl & Tobias Wolbring
- 171 Controlling for Taste Preferences – A Factorial Survey about
the Orientation to Judgment Devices in Movie Choice
Clemens Maria Schmidt
- 201 Fear of the Dark? A Systematic Comparison of Written
Vignettes and Photo Vignettes in a Factorial Survey
Experiment on Fear of Crime
Stefanie Eifler & Knut Petzold
- 235 The Impact of Presentation Format on Conjoint Designs:
A Replication and an Extension
Sophie Cassel, Josefine Magnusson & Sebastian Lundmark
- 273 Improving Anchoring Vignette Methodology in Health
Surveys with Image Vignettes
*Mengyao Hu, Sunghee Lee, Hongwei Xu,
Roberto Melipillán, Jacqui Smith & Arie Kapteyn*
- 315 New Methodical Findings on D-Efficient Factorial Survey
Designs: Impacts of Design Resolution on Aliasing and
Sample Size
Julia Kleinewiese
- 335 Factorial Surveys with Multiple Ratings per Vignette. A
Seemingly Unrelated Multilevel Regressions Framework
Alexander W. Schmidt-Catran
-

- 361 Information for Authors

Editorial: Vignette Analysis: Methodology and Recent Developments

*Lena M. Verneuer-Emre*¹, *Stefanie Eifler*² & *Hermann Dülmer*³

¹ RWTH Aachen University

² Catholic University of Eichstätt-Ingolstadt

³ University of Cologne and University of Passau

The term ‘vignette analysis’ draws on various disciplinary traditions to refer to various techniques for measuring normative judgements, subjective beliefs, and behavioural intentions on the basis of respondents’ answers to (a number of) brief descriptions of hypothetical situations, persons, or objects. The use of vignettes in survey research has been suggested within the framework of the ‘indirect measurement movement’ in empirical social research (Campbell, 1950) with the intention of bringing social context information into measurement. Numerous methodological studies have been undertaken with the aim of scrutinizing the assumed advantages of using vignettes.

The idea for this special issue was born in 2019, when we had the pleasure of hosting a session at the Conference of the *European Survey Research Association* (ESRA) in Zagreb, that brought together researchers with a special interest in research on vignette analyses. It was here that we again noticed the diversity of findings and the different ways of using vignette analyses, ranging from genuine methodological contributions through to applications of vignettes in the context of substantive research. A similar picture now emerges in this special issue: The contributions present methodological research on vignette analyses and innovative applications of this method, mostly located within the framework of experimental designs like factorial survey experiments, but also in the context of more general applications of vignettes such as anchoring vignettes or conjoint analyses.

The diversity of research findings on vignette analyses is our starting point in this editorial. The overall structure of this special issue of *mda* is as follows: The first chapter starts with a detailed literature review of factorial survey experiments to provide an overview of developments and trends in recent decades. In the same context, the second chapter provides an illustrative example for an application of factorial survey experiments. Subsequently, this special issue discusses two crucial

aspects relating to the application of vignettes – presentation and design resolution: Chapters three to five are dedicated to presentation format of vignettes in the context of factorial survey experiments, conjoint analyses, and anchoring vignettes. Chapter six focusses on design resolutions and the computer-based determination of the resolution IV design in *SAS On Demand for Academics*. The final chapter takes up the rarely used estimation technique of seemingly unrelated models in the context of factorial survey experiments. The papers collected and structured in this way in this special issue are framed within current research topics and findings more precisely below.

We start our framing of the collected papers with what is a truly outstanding contribution to the ‘indirect measurement movement’ already referred to, i.e., the factorial survey approach, which was introduced by Rossi (1979) as proposed by Paul F. Lazarsfeld (c.f., among others, Wallander, 2009). By transferring the basic principles of the *factorial* design (*multivariate* experimental design) into a sample *survey* (cf. Rossi & Anderson, 1982; Dülmer, 2007), the factorial survey combines both the high internal validity of causal inferences from experimental designs with the principally high external validity of causal inferences from survey research (Sniderman & Grob, 1996; Mutz, 2011; Auspurg & Hinz, 2015), regarding the generalizability of results to the broader population (cf. Sniderman & Grob, 1996; Auspurg & Hinz, 2015). Factorial surveys employ an experimental design that permits general conclusions to be drawn about causal mechanisms even without a random sample of respondents (cf. Auspurg & Hinz, 2015).¹

The factorial survey approach has been applied widely throughout the social sciences in recent decades. Studies have been undertaken on topics such as choosing the appropriate experimental design in factorial surveys (Atzmüller & Steiner, 2010; Dülmer, 2007; 2016), the effects of order, variation, wording, and presentation mode (Auspurg & Jäckle, 2017; Eifler & Petzold, 2014; Sauer et al., 2020; Shamon et al., 2022), choosing the most appropriate answer scale (Auspurg & Hinz, 2015; Sauer et al., 2020), learning and fatigue effects (Auspurg & Jäckle, 2017; Shamon et al., 2022), and the susceptibility of vignettes to social desirability response bias (Eifler, 2007; 2010; Eifler & Petzold, 2019; Groß & Börensen, 2009; Petzold & Eifler, 2020; Petzold & Wolbring, 2019). So far, the results of these studies are multifaceted and partly inconclusive, thereby giving rise to further questions.

1 While there are several approaches to the analysis of causal relationships with different research designs, many social scientists consider particularly the group of *experimental designs* as the *silver bullet* to the analysis of causal relationships (Shadish et al., 2002). The reason for this is that, in an experiment, social scientists “manipulate the presumed cause and observe the outcome afterward” (Shadish et al., 2002: 6) instead of considering social phenomena as they naturally occur in order to study causal relationships.

Edgar Treischl and *Tobias Wolbring* draw on the work of Lisa Wallander (2009) to open the special issue with a detailed literature review of factorial survey experiments published between 1982 and 2018. Besides looking at the development of research focussing on factorial survey experiments, the authors also focus on methodological advances as well as open questions in this research field. Their review shows that more and more research has been undertaken in this field over a period of several years, both with regard to attitude research and to issues relating to behavioural research topics. At the same time, the authors identify unresolved methodological challenges concerning the validity of vignettes and related realism issues.

As Treischl and Wolbring as well as the growing research on factorial surveys show, there are no substantive limits to the use of factorial designs. It is – as it is for all empirical analysis – mainly a question of the specific objective of the research and the appropriate implementation that may lead to the application of factorial designs. Generally speaking, the common denominator of factorial surveys is that they all aim to identify the relevant factors for judgements or behavioral intentions while studying social phenomena.

Clemens Maria Schmidt draws on Lucien Karpik's 'Economics of Singularities' to analyze the choice of movies using the Factorial Survey Approach. Due to the subjectivity of such a choice, the uncertainty of judgements is in Schmidt's view best anticipated by applying a factorial survey experiment in a student sample. As well as arriving at the interesting finding that diverse social devices are used to choose a film, Schmidt discusses the advantages of the factorial survey method and in particular how it supports analysis of the causal influence of those devices in situations where a choice has to be made.

Next, we consider the decisions researchers have to make when planning the application of vignettes in a survey: Besides the transformation of theoretical assumptions into situational descriptions, dimensions and levels to be depicted, challenges also arise with approximation to realism and the adequacy of the presented situation when applying vignettes in surveys. One crucial decision when setting up a vignette design concerns *the way vignettes are presented* to respondents. Vignettes were initially and, in most cases, continue to be presented as detailed written situational descriptions or in the form of short statements (e.g., Armacost et al., 1991; Triandis et al., 1998; Wallander, 2009). For some time now, studies have also used photos or videos (e.g., Golden III et al., 2001; Eifler, 2007; Noel et al., 2008; Krysan et al., 2009) as the presentation mode. First attempts have even been made to use virtual reality to present scenarios to respondents and in this way to focus on realism issues using immersive techniques (e.g., van Gelder et al., 2019). Whereas most applications make use of either written or visual vignettes, little research has so far been undertaken on the systematic comparison of different formats and their

effects. The findings that do exist are rarely clear-cut (Rashotte, 2003; Eifler, 2007; van Gelder et al., 2019). We are all the more pleased therefore to have three contributions that focus on presentation format with reference to very recent research contributions – in the context of factorial survey experiments, conjoint analyses, and anchoring vignettes:

Drawing on the theoretical perspectives of broken windows theory and the topic of fear of crime, *Stefanie Eifler* and *Knut Petzold* apply a split ballot experiment to compare different presentation formats of vignettes (written and photo) in a factorial survey. The authors investigate whether the context presented in a photo vignette leads to higher context approximation and thus to more valid answers than when using (classic) written vignettes. Overall, it is shown that the presentation format makes no difference to the assumed level of fear of crime of the vignette-dimensions. The presentation format was only observed to have an effect for setting characteristics (e.g., darkness) in the photo vignette.

Experiments that are closely related to vignette analysis are conjoint analysis (Luce & Turkey, 1964) and choice experiments (McFadden, 1974; cf. also Auspurg & Hinz, 2015). While the term “vignette analysis” prevails in social sciences, the term “conjoint analysis” traditionally dominates in marketing research where researchers are usually interested in the preference order for certain products. However, the basic structure of the experimental design for conjoint analysis and vignette analysis is the same, except that traditional conjoint analysis does not use confounded designs and all factors have to influence the judgement behaviour independently of each other (additive model without interaction terms, cf., Louviere, 1994). *Sophie Cassel*, *Josefine Magnusson* and *Sebastian Lundmark* focus on the presentation format in the context of such conjoint designs. The authors replicate the work of Shamon, Dülmer, and Giza (2019) and extend it to a paired conjoint experiment. Following a direct replication and analysis of the results of the extension, the authors confirm the conclusion that the table format is to be preferred to the text format in conjoint experimental designs.

Mengyao Hu, *Sunghye Lee*, *Hongwei Xu*, *Roberto Melipillán*, *Jacqui Smith*, and *Arie Kapteyn* contribute to the application of anchoring vignettes in health surveys with a special focus on the presentation format of these vignettes. In general, the challenge of inconsistent survey responses may arise due to diverse understandings of the subject in question – a problem that cannot be accounted for after data collection. The application of anchoring vignettes as an additional measurement tool in the process of data collection is one way of accounting for this difficulty: With the help of anchoring vignettes, the proportion of incomparability can be extracted in the process of analyzing the gathered data (cf. King et al., 2004; King & Wand, 2007; Hopkins & King, 2010; van Soest et al., 2011).

Hu et al. propose the use of image anchoring vignettes to overcome problems of complexity and time. By using data from a cross-cultural experiment and

comparing text and image vignettes, the authors conclude that image vignettes can improve respondents' differentiation of intensity levels, response consistency as well as the survey time in general.

Finally, D-efficiency in combination with design resolution and set size is also discussed. The higher the *D-efficiency* of a quota design, the lower the correlations between different vignette dimensions and the more balanced are the levels of each vignette dimension (Kuhfeld, 1997, cf. also Dülmer, 2016). The same applies to interaction terms, provided that they were included when a D-efficient design was generated. A design's *resolution* provides information about the aliasing (confounding) structure within a vignette set and/or about the confounding structure across the different quota sets selected by the researcher: the higher a design's resolution, the more main effects and higher order interaction effects, that are perfectly uncorrelated with other (higher order) interaction effects, can be estimated (McLean & Anderson, 1984; Ryan, 2007; Kuhfeld, 2010; cf. also Dülmer, 2016). Hence, higher design resolutions ensure a better protection from possible biases in the estimated effects than lower design resolutions. The disadvantage of a higher design resolution, however, is usually seen in the higher set sizes that are required for such designs.

Julia Kleinewiese contributes to the crucial topic of design resolution by focusing on quota designs, more precisely on D-efficient designs, and looks closely at the two-way interactions in resolution IV designs as well as the (minimum) number of vignettes (set size) for reaching a D-efficiency above 90 and as closely as possible to 100 (uncorrelated, balanced designs). Driven by the aim of an application-oriented paper, the author compares the aliasing structure of resolution IV designs as defined in the literature with the structure created by *SAS On Demand for Academics*. As well as discussing and reflecting on her finding of a discrepancy between the two, Kleinewiese also draws conclusions for the application of D-efficient designs and suggests, if possible, using resolution V designs as a standard design resolution in the social sciences.

Strategies of data analyses are of special interest for researchers who apply factorial surveys. By presenting several situational descriptions with varying dimensions to respondents, the data requires special treatment due to its hierarchical structure. Multilevel modelling is therefore the recommended choice for analyzing data with several ratings per respondent produced by factorial designs (Snijders & Boskers, 2012; Dülmer, 2016). A special case arises for factorial designs that are designed to measure not only different ratings per respondents but that also present several rating options for each vignette and thus produce multiple ratings per vignette.

Alexander Schmidt-Catran draws on this type of data structure and proposes an approach to statistically account for multiple ratings per vignettes with a Seem-

ingly Unrelated Regression framework. This approach – located within Structural Equation Modelling techniques – enables coefficients to be compared across ratings as well as the factor structure underlying such ratings to be analyzed. The author aims to make his proposal accessible to researchers by providing two application examples and the syntax in an online appendix.

References

- Alexander, C. S., & Becker, H. J. (1978). The Use of Vignettes in Survey Research. *Public Opinion Quarterly*, 42(1), 93-104.
- Armacost, R. L., Hosseini, J. C., Morris, S. A., & Rehbein, K. A. (1991). An Empirical Comparison of Direct Questioning, Scenario, and Randomized Response Methods for Obtaining Sensitive Business Information. *Decision Sciences*, 22(5), 1073-1090.
- Atzmüller, C., & Steiner, P. M. (2010). Experimental Vignette Studies in Survey Research. *Methodology*, 6(3), 128-138.
- Auspurg, K., & Hinz, T. (2015). *Factorial Survey Experiments*. Sage University Paper Series on Quantitative Applications in Social Sciences, 07-175. Los Angeles: Sage.
- Auspurg, K., & Jäckle, A. (2017). First Equals Most Important? Order Effects in Vignette-Based Measurement. *Sociological Methods & Research*, 46(3), 490-539.
- Campbell, D. T. (1950). The Indirect Assessment of Social Attitudes. *Psychological Bulletin*, 47, 15-38.
- Dülmer, H. (2007). Experimental Plans in Factorial Surveys: Random or Quota Design? *Sociological Methods & Research*, 35(3), 382-409.
- Dülmer, H. (2016). The Factorial Survey: Design Selection and its Impact on Reliability and Internal Validity. *Sociological Methods & Research*, 45(2), 304-347.
- Eifler, S. (2007). Evaluating the Validity of Self-Reported Deviant Behavior Using Vignette Analyses. *Quality & Quantity*, 41, 303-318.
- Eifler, S. (2010). Validity of a Factorial Survey Approach to the Analysis of Criminal Behavior. *Methodology*, 6(3), 139-146.
- Eifler, S., & Petzold, K. (2014). Der Einfluss der Ausführlichkeit von Vignetten auf die Erfassung prosozialer Einstellungen: Ergebnisse zweier Split-Ballot Experimente. *Soziale Welt*, 65(2), 247-269.
- Eifler, S., & Petzold, K. (2019). Validity Aspects of Vignette Experiments: Expected “What-If”-Differences between Reports of Behavioral Intentions and Actual Behaviour. In P. Lavrakas, M. Traugott, C. Kennedy, A. Holbrook, E. de Leeuw, & B. West (Eds.), *Experimental Methods in Survey Research: Techniques that Combine Random Sampling with Random Assignment* (pp 393-416). New York: Wiley.
- Golden III, J. H., Johnson, C. A., & Lopez, R. A. (2001). Sexual Harassment in the Workplace: Exploring the Effects of Attractiveness on Perception of Harassment. *Sex Roles*, 45(11/12), 767-784.
- Groß, J., & Börensen, C. (2009). Wie valide sind Verhaltensmessungen mittels Vignetten? Ein methodischer Vergleich von faktoriellem Survey und Verhaltensbeobachtung. In P. Kriwy & C. Gross (Eds.), *Klein aber fein! Quantitative Sozialforschung mit kleinen Fallzahlen* (pp 149-178). Wiesbaden: VS Verlag für Sozialwissenschaften.

- Hopkins, D. J., & King, G. (2010). Improving Anchoring Vignettes: Designing Surveys to Correct Interpersonal Incomparability. *Public Opinion Quarterly*, 74(2), 201-222.
- King, G., Murray, C. J. L., Salomon, J. A., & Tandon, A. (2004). Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research. *American Political Science Review*, 98, 191-207.
- King, G., & Wand, J. (2007). Comparing Incomparable Survey Responses: Evaluating and Selecting Anchoring Vignettes. *Political Analysis*, 15, 46-66.
- Krysan, M., Couper, M. P., Farley, R., & Forman, T. A. (2009). Does Race Matter in Neighborhood Preferences? Results from a Video Experiment. *American Journal of Sociology*, 115(2), 527-559.
- Kuhfeld, W. F. (1997). Efficient Experimental Designs Using Computerized Searches. In Sawtooth Software (Ed.), *Sawtooth Software*. Research Paper Series (pp 1-14). Retrieved July 13, 2022 (<https://homepage.stat.uiowa.edu/~gwoodwor/AdvancedDesign/KuhfeldTobiasGarratt.pdf>).
- Kuhfeld, W. F. (2010). Experimental Design: Efficiency, Coding, and Choice Designs. In W. F. Kuhfeld (Ed.), *Marketing Research Methods in SAS. Experimental Design, Choice, Conjoint, and Graphical Techniques* (pp 53-241). Retrieved July 13, 2022 (<http://support.sas.com/techsup/technote/mr2010c.pdf>).
- Louviere, J. J. (1994). Conjoint Analysis. In R. P. Bagozzi (Ed.), *Advanced Methods of Marketing Research* (pp 223-259). Cambridge, MA: Blackwell Publishers.
- Luce, R. D., & Tukey J. W. (1964). Simultaneous Conjoint Measurement: A New Type of Fundamental Measurement. *Journal of Mathematical Psychology*, 1, 1-27.
- McFadden, D. (1974). Conditional Logit Analysis of Qualitative Choice Behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics* (pp 105-142). New York, NY: Academic Press.
- McLean, R. A. & Anderson, V. A. (1984). *Applied Factorial and Fractional Designs*. New York, NY: Marcel Dekker.
- Mutz, D. (2011). *Population-Based Survey Experiments*. Princeton: Princeton University Press.
- Noel, N. E., Maisto, S. A., Johnson, J. D., Goings, C. D., & Hagman, B. T. (2008). Development and Validation of Videotaped Scenarios: A Method for Targeting Specific Participant Groups. *Journal of Interpersonal Violence*, 23(4), 419-436.
- Petzold, K., & Eifler, S. (2020). Die Messung der Durchsetzung informeller Normen im Vignetten- und Feldexperiment. In I. Krumpal, & R. Berger (Eds.), *Devianz und Subkulturen* (pp 167-204). Kriminalität und Gesellschaft. Springer VS: Wiesbaden.
- Petzold, K., & Wolbring, T. (2019). What Can We Learn from Factorial Surveys About Human Behavior? A Validation Study Comparing Field and Survey Experiments on Discrimination. *Methodology*, 15, 19-30.
- Rashotte, L. S. (2003). Written Versus Visual Stimuli in the Study of Impression Formation. *Social Science Research*, 32(2), 278-293.
- Rossi, P. H. (1979). Vignette Analysis: Uncovering the Normative Structure of Complex Judgments. In: R. K. Merton, J. S. Coleman, & P. P. Rossi (Eds.), *Qualitative and Quantitative Social Research. Papers in Honor of Paul F. Lazarsfeld* (pp 176-186). New York: The Free Press.

- Rossi, P. H., & Anderson, A. B. (1982). The Factorial Survey Approach: an Introduction. In P. H. Rossi, & S. L. Nock (Eds.). *Measuring Social Judgments. The Factorial Survey Approach* (pp 15-67). Beverly Hills, CA: Sage.
- Ryan, Th. P. (2007). *Modern Experimental Design*. Hoboken, NJ: John Wiley & Sons.
- Sauer, C., Auspurg K., & Hinz T. (2020). Designing Multi-Factorial Survey Experiments: Effects of Presentation Style (Text or Table), Answering Scales, and Vignette Order. *Methods, Data, Analyses, 14*(2), 195-214.
- Shadish, W. R., Cook, T. D., & Campbell, D. D. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Belmont, CA: Wadsworth.
- Shamon, H., Dülmer, H., & Giza, A. (2022). The Factorial Survey: The Impact of the Presentation Format of Vignettes on Answer Behavior and Processing Time. *Sociological Methods & Research, 51*(1), 396–438.
- Sniderman, P. M., & Grob, D. B. (1996). Innovations in Experimental Design in Attitude Surveys. *Annual Review of Sociology, 22*, 377-399.
- Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modeling*. Second Edition. Los Angeles: Sage.
- Triandis, H. C., Chen, X. P., & Chan, D. K.-S. (1998). Scenarios for the Measurement of Collectivism and Individualism. *Journal of Cross-Cultural Psychology, 29*(2), 275–289.
- van Gelder, J.-L., de Vries, R. M., Demetriou, A., van Sintemartensdijk, I., & Donker, T. (2019). The Virtual Reality Scenario Methods: Moving from Imagination to Immersion in Criminal Decision-Making Research. *Journal of Research in Crime and Delinquency, 56*(3), 451-480.
- Van Soest, A., Delaney, L., Harmon, C., Kapteyn, A., & Smith, J. P. (2011). Validating the Use of Anchoring Vignettes for the Correction of Response Scale Differences in Subjective Questions. *Journal of the Royal Statistical Society. Series A: Statistics in Society, 174*, 575-595.
- Wallander, L. (2009). 25 Years of Factorial Surveys in Sociology: A Review. *Social Science Research, 38*, 505-520.

The Past, Present and Future of Factorial Survey Experiments: A Review for the Social Sciences

Edgar Treischl & Tobias Wolbring

Friedrich-Alexander-University

Abstract

Factorial survey experiments (FSEs) are increasingly used in the social sciences. This paper provides a review about the use of FSEs and aims to answer three research questions. (1) How has this specific research field developed over time? (2) Which methodological advances have been made in FSE research and to what degree are they applied in empirical studies? (3) Which questions remain unresolved and should be addressed in future research? Using the Web of Science and Scopus databases, we conducted a literature review of FSEs published between 1982 and 2018. Our findings show that the field is developing quickly and that FSEs are becoming increasingly accepted in different research areas. Thereby, FSEs are being widely used not only to study attitudes, but also to explore the determinants of behaviour. Most research applies state-of-the-art techniques in terms of statistical analysis; however, to a lesser extent, studies rely on more sophisticated sampling procedures to draw samples from a large vignette universe. Finally, several methodological questions remain unresolved concerning the realism and complexity of vignettes, social desirability, and the predictive validity of FSEs regarding behaviour due to their hypothetical nature. Against this background, we call for more methodological research to assess the general applicability of FSEs for different research areas. Further, our review suggests the need for better documentation and reporting standards to evaluate methodological aspects of FSEs.

Keywords: factorial survey experiments, methodological advances and pitfalls, predictive validity, realism of vignettes, vignette design



In 2009, Lisa Wallander published a highly cited review article about factorial survey experiments (FSEs). As she pointed out, many scholars were not familiar with them or had substantial reservations against them at the time, even though they had been introduced over three decades prior (see Jasso & Rossi, 1977; Rossi et al., 1974; Sampson & Rossi, 1975). As a result, empirical studies using FSEs were scarce. Figure 1 displays the number of articles published between 1982 and 2018 that refer to an FSE and have been identified in our review. As Figure 1 shows, only a few papers using FSEs were published every year until 2006, which was the last year covered in Wallander's review. Further, FSEs were virtually absent in leading social science journals.

A decade later, the situation has changed: FSEs have been introduced into survey methodological handbooks (see Aviram, 2012), textbooks are available that explain how to design and conduct FSEs in detail (see Auspurg & Hinz, 2015a; Mutz, 2011), and multifactorial survey experiments are becoming increasingly popular in the social sciences (see Atzmüller & Steiner, 2010; Auspurg & Hinz, 2015b; Jasso, 2006). In accordance with this trend, the number of publications using FSEs has risen markedly since 2006, as Figure 1 indicates. Several of these studies were published in leading journals, such as the *American Sociological Review* and the *European Sociological Review* (e.g. Auspurg et al., 2017; Graeff et al., 2014; Wouters & Walgrave, 2017), which further illustrates the increasing use and acceptance of FSEs.

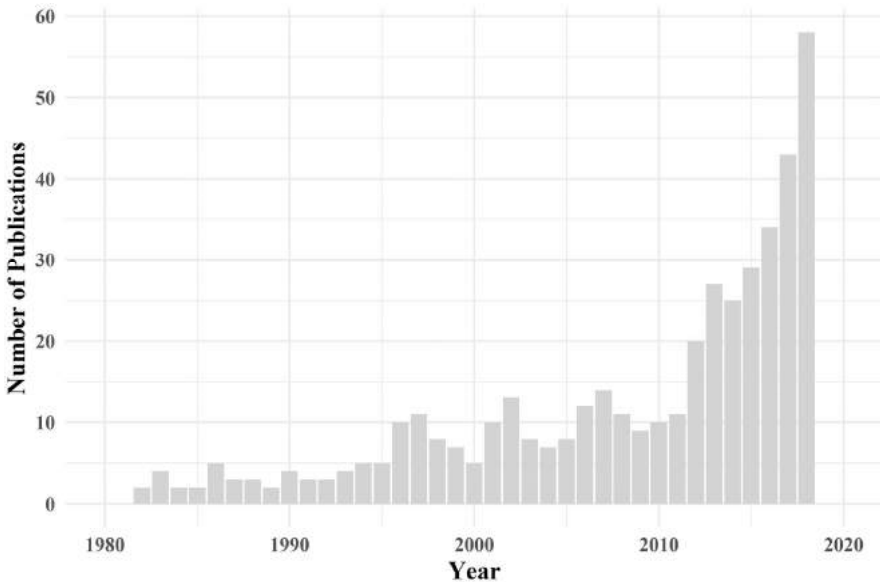
Given the popularity of FSEs and the increasing number of empirical applications since Wallander published her influential review article, we think it is time for an update. Hence, we focus on how the field has developed, which methodological advances have been made, and which challenges of the approach remain unresolved. For this reason, we conducted a literature review covering all articles from Wallander's review article (1982–2006), as well as more recent applications involving FSEs (2007–2018). For each publication, we collected information about the study topic, research design, outcome measures, and statistical analysis. This gives us the opportunity to make three contributions to the current literature on FSEs. First, we provide an overview of the past and current use of FSE in the social

Acknowledgements

This paper could not have been realized without the support of Janette Buchmann, Sarah Glaab, Nora Spielmann, and Philipp Überall during the data collection process. For helpful feedback, we want to thank Johanna Gereke, Eva Zschirnt, the anonymous reviewers, the editors as well as all participants at the European Sociological Association midterm conference 2018 in Cracow.

Direct correspondence to

Edgar Treischl, Friedrich-Alexander-University, Erlangen-Nürnberg (FAU),
Findelgasse 7/9, 90402 Nürnberg, Germany
E-mail: edgar.treischl@fau.de



Note: Number of articles published between 1982 and 2018 that refer to an FSE and have been identified in our review. For details on the literature review, see Section 3.

Figure 1 Number of published FSE articles

sciences. We identify research areas in which the approach is increasingly applied. We have also updated Wallander's review regarding some basic methodological choices such as sampling strategies, the respondents' countries of origin, and between- versus within-subjects designs.¹

Second, since some recommendations on how to design and analyse an FSE have been published in recent decades, we briefly introduce readers to these methodological advances, and we examine to what extent these techniques have entered applied research. Methodological advances can help to improve both the internal validity of inferences and the statistical power. Thus, one goal of our review is to

1 Choice experiments are another kind of multifactorial survey experiment. In choice experiments, participants are directly confronted with varying trade-offs between two or more alternatives, and are asked to choose between the proposed alternatives. Choice experiments appear to be especially well-suited to studying human decision-making since they are theoretically grounded in the characteristics theory of value (Lancaster, 1966) and random utility models (McFadden, 1974; Manski, 1977). In this review, we focus on FSEs as the most widely used type of multifactorial survey experiment in the social sciences, while choice experiments are more frequently employed in business studies and economics (for the potentials and challenges of choice experiments in the social sciences see Liebe & Meyerhoff, 2021).

give researchers some general background on how to design state-of-the-art FSEs and to provide references for more detailed follow-up.

Finally, we also aim to provide guidance for future methodological research by highlighting unresolved questions in the growing, but small, methodological literature about FSEs. We focus on three partly interrelated issues that have caused controversial discussions within the scientific community, as they may have far-reaching consequences for the validity of FSEs: the realism and complexity of vignettes, concerns regarding the hypothetical nature of the outcome measures in FSEs, and the risk of social desirability bias. While methodological research on these topics is still scarce, we underscore some findings from recent research about the design of FSEs, and contrast these methodological recommendations and insights with current research practices as identified by our literature review.

The remainder of the paper is structured as follows: First, we introduce the FSE approach and provide some methodological background on it. Next, we outline in more detail how we conducted the literature review and describe the dataset. Furthermore, we explain some recent methodological advances and discuss unresolved questions such as the required degree of realism and complexity of vignettes, as well as the link between stated and actual behaviour. Finally, we emphasise key insights and opportunities for future research to deepen our methodological knowledge of FSEs.

The Basic Idea Behind Factorial Survey Experiments

This section outlines the basic idea behind FSEs.² Respondents encounter textual descriptions or visual stimuli of a hypothetical situation (*vignette or scenario*) in an FSE and are asked to rate the scenario. Each vignette contains one or several characteristics (*dimensions/factors*) that systematically vary across vignettes. Survey participants are randomly assigned to one (*between-subjects design*) or several (*within-subjects design*) vignettes, and are asked for their opinion on a certain situation or the intended behaviour in the described scenario.

Figure 2 displays two examples of vignettes. Example A is a vignette by Opp (2002). He examined under which circumstances an anti-smoking norm emerges by eliciting normative judgements. Single dimensions that may have a causal impact on respondents' opinions are in *italics* to illustrate the experimental variation across the vignettes. Example B comes from a study by Teti et al. (2016). They

2 Several excellent textbooks about FSEs have been published (see Auspurg & Hinz, 2015a; Mutz, 2011) since the approach was first introduced to the social sciences by Rossi et al. in 1974. This section relies heavily on these textbooks, which provide a more detailed discussion about the fundamentals of FSEs.

Example A (Opp, 2002):

Mr. Müller goes to a restaurant. This is a *top class restaurant* in which smoking is prohibited. There is *nobody* in the restaurant who smokes. Mr. Müller stays *only for a short time* to drink a beer. He smokes most of the time, *more than a package of cigarettes per day*.

Example B (Teti et al., 2016):

Imagine that the apartment offered is in *your current district*. It is located very *centrally*, *2 minute walk* from the nearest bus/train station and *far away* from the home of your daughter/son. The apartment is in the 3rd floor, *has no elevator*, and has a *large bathtub* (no shower) and a *balcony without steps*.

Figure 2 Two examples of vignettes

asked elderly respondents to make hypothetical relocation decisions and investigated whether FSEs can be applied in housing research.

An FSE combines the methodological rigour of an experimental design with the advantages of survey research by including an experimental research module in a survey and assigning participants *randomly* to one or several hypothetical descriptions of a situation. This facilitates inferences from experimental results to a target population (see Auspurg & Hinz, 2015a, p. 12-13; Mutz, 2011, p. 10).

Observational studies may suffer from various methodological pitfalls—such as confounding by self-selection of participants and unobserved heterogeneity—thus impairing the identification of causal effects (Rosenbaum, 2010; Shadish et al., 2002). As is well-known, experimental designs have advantages regarding causal inference and, at least in theory, can outperform non- or quasi-experimental designs in regard to issues of internal validity (for the principles of experimental design, see Imbens & Rubin, 2015; Jackson & Cox, 2013). An FSE offers the possibility of estimating the causal effect of a varied dimension on the outcome variable. Random assignment helps to avoid threats to internal validity such as confounding and selection bias. Direct manipulation of treatments (in the FSE, the varied dimension) secures causal ordering, and including a control group avoids biases due to maturation effects and study participation.

Furthermore, the survey implementation of the FSE helps to address problems common in experimental research. In particular, lab experiments are often criticised for a lack of external validity and transportability, since they rely on participants (mostly students) from Western, educated, industrialised, rich and democratic ('weird') societies (Bader et al., 2019; Henrich et al., 2010). In a similar vein, not only lab, but also field experiments are often challenged by the infeasibility of randomised trials due to ethical concerns, practical restrictions, and lack of manipulability of the treatment (Deaton & Cartwright, 2018; Teele, 2014). Such problems

can be avoided by using textual descriptions of hypothetical scenarios instead of actual interventions in the ‘real’ world.

It is easier to sample non-students and ‘non-weird’ people for a survey than to recruit them for lab experiments. Including an experimental module in the survey allows scholars to conduct a population-based FSE, promising broader generalisability beyond potentially selective subgroups such as students. Variations are much easier to implement in an FSE due to the manipulation of textual descriptions. Ethical concerns and practical restrictions do not apply to the same degree as in the lab or in the field. Accordingly, treatments that are hard to implement in the field can be investigated in an FSE. For this reason, an FSE can also help to inform policy about hypothetical worlds and potential interventions discussed in the public discourse without taking the risk and covering the costs of an actual implementation. As the following review shows, FSEs are increasingly being used in the social sciences, even though FSEs also face methodological pitfalls and challenges.

Literature Review

After this short introduction to FSE, this section informs about the literature review in two sub-sections. In the first sub-section, we provide details about the data collection process, search strategy, and inclusion restrictions for the literature review. In the second sub-section, we update Wallander’s review by describing our analytical sample in terms of research areas (e.g. topics, the respondents’ countries of origin) and methodological choices (e.g. sampling strategies, between- vs. within-subjects designs).

Data Collection

Our literature review is based on a combination of three different approaches to secure broad coverage of FSE publications. First, we covered all 106 publications that Wallander (2009) identified. Second, we made use of the popularity of the first review paper and collected publications citing it. In 2019, Wallander had over 400 citations according to Google Scholar, including many recent FSE applications. Third, we searched for empirical applications of FSEs using the Web of Science and the Scopus database for the time period covered by Wallander (1982–2006), as well as more recent years (2007–2018).³

3 For identifying relevant publications among papers citing Wallander (2009), we applied the same search strategy and criteria as outlined in the third search strategy. However, among those publications many appeared as monographs or were grey literature (such as working papers, project reports, and presentation slides). Consistent with the third search strategy, we did not include them in the review.

We applied the following restrictions to identify publications relevant for our review. The review included publications that refer to ‘factorial survey (experiments)’, ‘vignette study’ and ‘vignette experiment’ in the title, abstract, or keywords. Hence, the review covers FSEs, but not publications with related but different survey experimental research designs, such as conjoint analysis and discrete choice experiments. To identify core articles for the social sciences, we considered publications published in *journals* listed in the *Social Sciences Citation Index (SSCI)* of the Web of Science and the category ‘*Social Sciences*’ of Scopus. We did not cover other document types such as monographs, or conference articles. Further, we only took publications written in English into account. The following search string was used to identify FSEs using the Web of Science:⁴

TOPIC: (“Factorial survey”) *OR* **TOPIC:** (“Factorial survey experiment”) *OR* **TOPIC:** (“Vignette study”) *OR* **TOPIC:** (“Vignette experiment”) **Refined by:** [excluding] **PUBLICATION YEARS:** (1974 – 1981 *OR* 2019 *OR* 2020 *OR* 2021) *AND* **DOCUMENT TYPES:** (ARTICLE) *AND* **LANGUAGES:** (ENGLISH) *AND* **WEB OF SCIENCE INDEX:** (WOS.SSCI) **Timespan:** All years. **Indexes:** SCI-EXPANDED.

Similarly, the following search string was used to identify FSEs with Scopus:

(**TITLE-ABS-KEY** ((“factorial survey experiment”))) *OR* **TITLE-ABS-KEY** ((“factorial survey”)) *OR* **TITLE-ABS-KEY** ((“vignette study”)) *OR* **TITLE-ABS-KEY** ((“vignette experiment”))) *AND* **PUBYEAR** > 1981 *AND* **PUBYEAR** < 2019 *AND* (**LIMIT-TO** (SUBJAREA , “SOCI”)) *AND* (**LIMIT-TO** (DOCTYPE , “ar”)) *AND* (**LIMIT-TO** (SRCTYPE , “j”)) *AND* (**LIMIT-TO** (LANGUAGE , “English”))

Based on those search strings, we identified 148 publications in the Web of Science and 301 publications in Scopus for the entire time period (1982–2018; last search date: 26 March 2021), with substantial overlap between the two databases. After taking into account the overlap, 353 publications remain in the sample from the Web of Science and Scopus.

Upon closer inspection it turned out that not all of these 353 publications meet the scope condition of our review to report on empirical applications of FSE in the social sciences. Different reasons lead to the exclusion of some publications. The applied exclusion restrictions were as explained in the following (for the excluded number of publications by criteria see Table A1 in the online appendix). Further

4 We used the displayed search string to collect data from the Web of Science last time in March 2021. After the last search, the Web of Science database received various substantial updates and extensions in 2021, including a fundamental update of the search tool and a new code structure to search for publications. As a result, the reported search string of our review is not working with the recent version of the Web of Science.

inspection of the publications showed, that some publications did not employ an FSE but introduce the FSE methodology (e.g. Taylor, 2005). In a similar vein, some publications report only results of a pilot study to introduce FSEs or discuss them in light of a specific research area (e.g. Liebig et al., 2015). In some instances, qualitative researchers use vignettes and many health-related research use case scenarios (Kiesewetter et al., 2018) to describe a scenario, but without applying typical aspects of FSEs (e.g. random assignment, varying dimensions). In addition, past research sometimes confounds FSEs with factorial experiments (e.g. Baker, 1983). We did not include these in total 85 studies in our review.

In addition, a small but growing number of publications address methodological research questions on FSEs (e.g. for varying the number of vignette dimensions, see Auspurg & Jäckle, 2017). In the following, we will only provide statistics on substantive research using FSEs, and exclude review articles, as well as methodological research on FSEs. However, these contributions are part of our discussion on methodological advances. In this part of the review, we also discuss insights from more recent methodological contributions.

Finally, we decided to focus on studies with textual vignette descriptions (including tables), but did not include studies using visual stimuli, such as pictures and video vignettes in our review (e.g. see Oberoi et al., 2016; Wouters & Walgrave, 2017). The main reason was that these studies are often not completely comparable with research using text vignettes: Studies using video or photo vignettes often vary a lower number of dimensions due to the effort involved in manipulating visual stimuli, and the cognitive processes of the respondents when seeing visual stimuli might be fundamentally different from those when reading text.

Hence, our final dataset based on Scopus and the Web of Science contains 261 publications that met the described criteria. Moreover, the final data considers all 106 publications identified by Wallander (2009) and 74 publications that cite Wallander (2009) and were not listed in the Web of Science or in Scopus. Overall, our final analytical sample contains 441 publications. A list of included publications can be found in Table A2 in the online appendix.

We then created a dataset containing detailed information on each publication. We retrieved most information from the 'data and methods' section and, in some instances, from the 'appendix'. To summarise how the field has developed, we collected information about the research topic and classified the outcome measurement of each publication to indicate whether respondents were asked to make a hypothetical judgement (e.g. fairness of earnings) or to state a behavioural intention (e.g. willingness to pay for a service). In addition, the data contain information about the survey sampling strategy, the number of vignette dimensions, the measurement of the outcome, the vignette sampling strategy, and the applied statistical analysis.

Unfortunately, some publications did not report details about the FSEs in terms of design. In particular, several recent publications did not contain information about how vignettes have been sampled from the vignette universe. The fact that we could not retrieve this information, even after an extensive search, is alarming and calls for establishing standards of how to document and report design aspects of FSEs.

All publications were classified by three different coders. The interrater reliability between the three raters was sufficient ($r=0.89$) for numerical indicators such as the number of dimensions, the number of ratings, or the outcome measure. In contrast, we found the lowest interrater reliability ($r=0.54$) for the binary indicator for sensitive research topics. We have not given detailed statistics on sensitivity in the paper, but cover the topic in our discussion on the predictive validity of FSEs.

Description of the Sample: Research Areas

Before focusing on methodological advances and unresolved concerns, we *update* the work of Wallander (2009). Figure 3 plots the top 10 FSE publication topics before and since 2007 in terms of absolute and relative frequency. We identified research areas in which FSEs have been increasingly used since Wallander's review, which is why we centre on 2007 as the cut-off point and examine the development of FSE publications before and since 2007. As classification is sometimes not straightforward (e.g. research on school-to-work transitions), the categories of the classification are not disjunct. An article could fall in two or more categories.⁵

As Figure 3 shows, most studies ($N=58$) have used FSEs to study *crime and justice* topics (Lyons, 2008; Tolsma et al., 2012). This research area was the most prevalent topic among FSE publications until 2006. The number of published FSEs about justice has decreased to 43 since 2007, but FSEs are still often applied in this field. In contrast, FSEs are increasingly applied in other areas over time. The categories *health and care* and *work* display the highest increase in absolute numbers, with 68 and 42 applications since 2007. Overall, 37% of all published FSEs that we identified examine *health and care*-related topics, such as care planning and needs (Baughman et al., 2019; Jörg et al., 2006), or *work*-related topics, such as hiring intentions (Di Stasio & Gërkhani, 2015; van Belle et al., 2018). However, research is not restricted to these topics. FSEs are used to study diverse aspects, and, as a result, we did not classify a certain number of studies under a separate category, but rather as *other topics* (overall 9%). This includes research about sport behaviour (Chatfield et al., 2018), corruption (Graeff et al., 2014), and the willingness to

5 For instance, Haase et al. (2016) examined the male breadwinning model, a topic that might be included in the work or family category. In such an instance, we included the article in both categories.

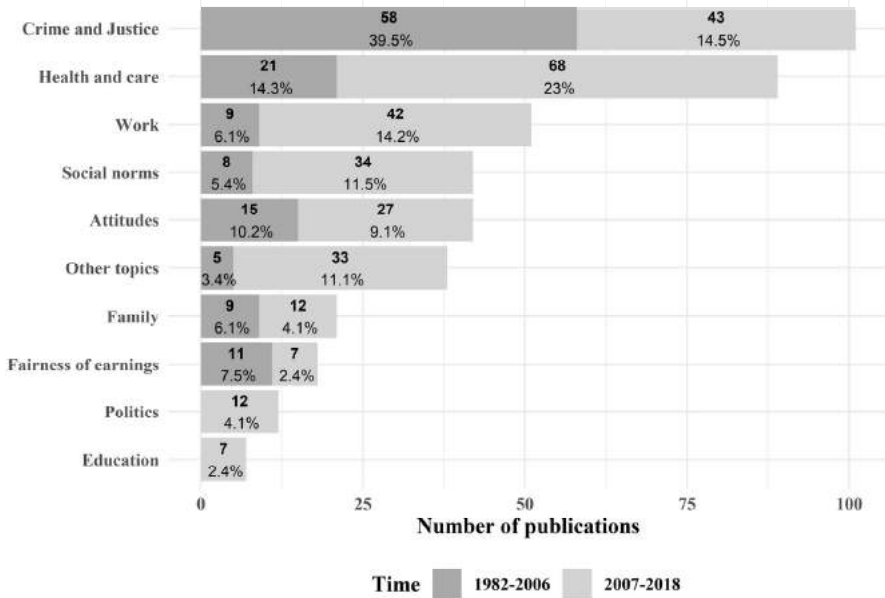


Figure 3 Top 10 FSE publication topics before and since 2007

provide (para) data (Couper & Singer, 2012). In addition to the top 10 topics, FSEs have less frequently been used to study conflict behaviour (see Baron et al., 2001; Bell & Forde, 1999), consumption (Moynihan, 2013; Cahan, 1996), and mobility behaviour (see Abraham et al., 2010; Teti et al., 2016). We classified, but excluded these topics in Figure 3 due to the small number of publications since 2007.⁶

With regard to respondents' countries of origin, Wallander (2009) reported studies from seven different countries, but over 80% were based on populations from the US. For our review, we observed 294 articles from 41 different countries since 2007. Many studies rely on US populations (31%), but a substantial number of publications come from other countries, chief among them the Netherlands (14%), Germany (13%), and the UK (7%). The growing number of respondents' countries of origin also illustrates the increased popularity.

6 In recent years, the amount of methodological research has grown as well, but methodological research remains rare in comparison to hot topics and the overall amount of FSE studies.

Description of the Sample: Applied Methods

In addition to the diversity of FSE regarding research topics and respondent's country of origin, Wallander (2009) reported that almost every second study aims to make inferences from experimental results to a general population. However, in the first review, it remained unclear which sampling strategy most researchers used for such inferences from the sample to apply to the target population. In the case of a non-probability sample, experiments still provide internally valid estimates of a treatment effect due to the random assignment of subjects to treatment conditions. However, the effect estimates of a convenience sample cannot necessarily be generalised beyond the specific subgroup under investigation in the case of effect heterogeneity across individuals. As our data show, approximately 47% use probability and 53% non-probability samples. Within the group of studies using non-probability samples, most authors have relied on convenience samples, in particular from the student body. However, in a few cases, researchers used referral (4%) and purposive samples (2%). Reflecting most recent studies, another 9% of non-probability samples have used samples from the crowdsourcing website Amazon Mechanical Turk (M-Turk). Hence, while the use of non-probability sampling might often be unproblematic for generalising experimental results if one is willing to assume the absence of effect heterogeneity, most FSEs do not fully utilise the potential of FSEs to generate 'representative' samples. As a consequence, in the most extreme case, the findings from a part of the literature might not be generalisable to the target population. In addition, the use of non-probability samples raises questions about the adequacy of inferential statistics, which is frequently applied with such data.

As Wallander revealed in her review, many studies have not applied methods for clustered data, even though a substantial number of vignette studies depend on a within-subjects design with two or more vignette ratings per respondent. In our review, 86% of the studies relied on a within-subjects design. The average number of vignettes per person is nine, with a maximum of 110, and 50% of the studies ask for five or more ratings. Given the broad use of within-subjects designs, the hierarchical structure of the data needs to be considered: Each person provides several ratings. Consequently, single observations are clustered and are no longer independent from each other. Clustered data violate the assumption of regression analysis that residuals are independent and identically distributed. Without adjustments for the clustered data structure, standard errors from a regression analysis are biased. The two most common ways to address this problem are (a) multilevel models with random or fixed effects and (b) robust standard errors clustered around individuals (see Maas & Hox, 2004; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012). Figure 4 contains a proportional stacked area chart to display the proportions of statistical methods used to analyse FSEs with a within-design over time. Forty-six percent of published articles in 2000–2004 presented the results of a regres-

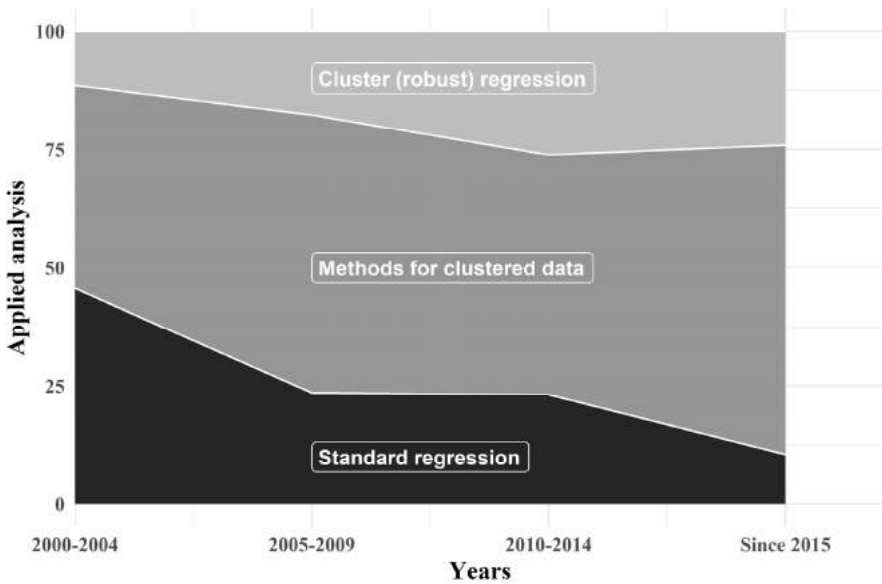


Figure 4 Statistical analysis methods in FSE publications

sion analysis without taking clustered data structure into account. This proportion fell considerably over time. Most studies published since the first review paper no longer ignore the issue. In the period of 2015–2018, the majority of most recent publications used methods for hierarchical data (69%) or relied on cluster-robust standard errors (26%), while only 5% did not take clustering into account.

Methodological Advances and Unresolved Questions

This section discusses more recent methodological insights and advances in the design of FSEs. First, we introduce ways to improve the efficiency of the vignette design in the face of a large vignette universe. Second, we provide recent methodological findings regarding the consequences of the vignette design for the validity of the results. We place particular emphasis on the complexity and realism of vignettes, but also cover issues of presentation style and choice of response scales. Finally, recent scholarly publications have examined the relationship between behavioural intentions stated in FSEs and actual behaviour. We discuss why FSEs may or may not help to provide insights into human behaviour, and provide an over-

view of existing studies examining the predictive validity of FSEs. While not all these methodological discussions will lead to clear recommendations, we believe it is important to draw attention to these topics, both for a reflected use of FSEs by applied researchers, and to provide motivation for future methodological research.

Design Efficiency

The *vignette universe* contains all vignettes, which result from the combination of all levels of each dimension (*full factorial*) in an FSE. For example, the vignette universe of an FSE with seven dimensions and four levels of each dimension contains $4^7=16.384$ unique vignettes as a result of the Cartesian product. Researchers can use the full factorial in the case of a small universe.⁷ However, the vignette universe quickly becomes very large. In such an instance, scholars must construct an experimental design such as *random sampling*, *randomised block confounded factorial (RBCF) designs*, and *D-optimal designs* to draw a smaller, more manageable subset from the vignette universe.

Random sampling techniques reduce the number of vignettes by drawing for each respondent a random set of vignettes from the universe. Random sampling techniques generate an orthogonal (FSE dimensions are uncorrelated) and completely unconfounded vignette set if the vignette sample approaches infinity, but not necessarily for smaller vignette sample sizes (Jasso, 2006; Su & Steiner, 2020). Both other *fractional factorial designs* try to actively increase the efficiency of the experimental plan compared to random sampling techniques.⁸ For instance, RBCF designs use experimental plans to split the vignettes into several vignette sets of equal size, such that only higher-order interaction effects are confounded with the sets.⁹ In a similar manner, a D-optimal design is the outcome of a computational optimisation process. A D-optimal design tries to maximise the precision of parameter estimates by searching for an orthogonal and balanced (levels have

7 This illustrates another advantage of FSEs. The levels of different dimensions are often highly correlated with each other in surveys and other observational studies (see Auspurg & Hinz, 2015a, p. 10). By using the full factorial, all considered dimensions in an FSE are uncorrelated by design (orthogonal) due to the combination of all dimensions and levels. Orthogonality is a main strength of FSEs since the causal effect of each dimension is identified in such a design.

8 Design efficiency refers to the statistical power of the vignette sample (experimental design) to estimate parameters for main dimensions and interaction effects with a high degree of precision. For more information about design efficiency and recent developments, see Dülmer (2016) or Su and Steiner (2020).

9 As Su and Steiner (2020, p. 36) denoted: “RBCF designs are typically restricted to simple designs with a few factors and ideally the same number of factor levels. For more complex designs that involve large vignette populations, generated from a large number of factors (i.e. five or more factors) with unequal numbers of factor levels (i.e. 2–10 or more levels), adequate RBCF designs might not exist or be challenging to construct”.

equal frequencies) vignette sample of the universe (see Dülmer, 2016). A computer algorithm searches iteratively for combinations of each dimension to optimise the precision of parameter estimates for all main effects and may—depending on specifications—also optimise precision for two-way or higher-order interactions (see Kuhfeld et al., 1994).

Thus, such *fractional factorial designs* have advantages compared to random samples. Researchers do not have to rely on chance that the random sample is the most efficient design and that key assumptions such as orthogonality hold. For instance, research indicates that D-optimal designs outperform random sampling techniques and a full factorial in the case of a small random sample from the vignette universe due to higher statistical power to estimate interaction effects (Dülmer, 2007). Especially in the social sciences, interaction effects are often of major interest. A random sample is not ideal to estimate these effects, and sometimes interactions are not identified by the design at all.

However, most past research has used random samples of the vignette universe. Only one study (Buskens & Weesie, 2000) out of 44 articles that relied on a vignette sample also used a fractional factorial design until 2006 (see Wallander, 2009, p. 512). In addition to the fact that those designs were not very well-known back then, most statistical software packages had not implemented packages at that time to draw a D-optimal design and to calculate a design's efficiency. Although fractional factorial designs are now broadly accepted as a useful sampling technique, the unfortunate situation on the software side remains almost unchanged (for an implementation in SAS, see Kuhfeld et al., 1994). Hence, we suspect that an increasing, but still minor share of recent studies uses a D-efficient design.

Our literature review corroborates this apprehension. Figure 5 depicts the use of different vignette sampling techniques over time based on a proportional stacked area chart. Random vignette samples were the most common technique in the period of 2000-2004. Seventy-one percent of all FSE articles report using a random sample of the vignette universe. As Figure 5 shows, there is a clear time trend. While the large majority of studies published during 2000–2004 used random samples of vignettes, only 25% of the identified publications after 2014 did so. However, random sampling techniques are far from being fully replaced by fractional factorial or full factorial designs, although we find an increasing amount of both, especially for full factorial designs since 2000. Unfortunately, a small, but growing amount of research does not provide any information about the process of vignette sampling.

In sum, random sampling techniques are easy to implement and might be a sufficient choice in the case of a small vignette universe and large sample sizes. However, random sampling techniques come with the risk of potentially confounding main and interaction effects, and require untestable assumptions about the absence of certain interactions. RBCF and D-optimal designs help to avoid

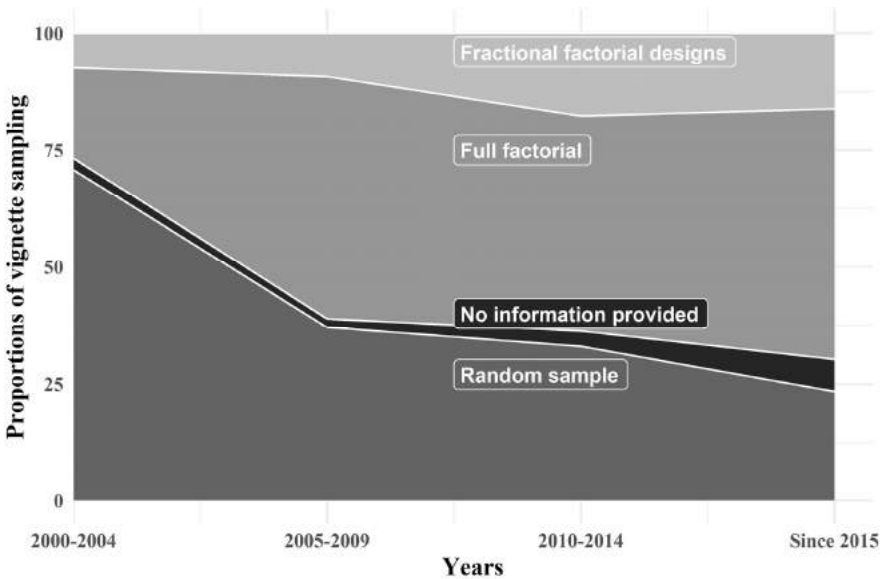


Figure 5 Vignette sampling strategies in FSE publications

such threats to internal validity by securing the orthogonality of main and interaction effects, and often increasing statistical power. Even if more time investment is needed to determine and implement these designs, we especially recommend using them in the case of small samples and a large vignette universe to avoid confounding and underpowered FSEs.

The Realism and Complexity of Vignette Designs

Decisions about the design of a vignette can have far-reaching consequences in terms of internal and external validity. In terms of *complexity*, a very simple scenario with only a few dimensions and rather low variation across several presented vignettes may lead, on the one hand, to boredom and fatigue effects in within-subjects designs. On the other hand, very detailed scenarios with many dimensions may seem more *realistic*, but providing too much information may cause cognitive overload, especially if the number of ratings is high. Participants may no longer be able or willing to pay attention to the vignette or to all provided dimensions in the case of information overload. Instead, participants may switch to response sets, use cues and heuristics to come to a decision without too much cognitive effort. Such satisficing behaviour is well-known for conventional survey items (see Krosnick, 1991) and can also occur in different forms in FSEs (see Shamon et al., 2019). For

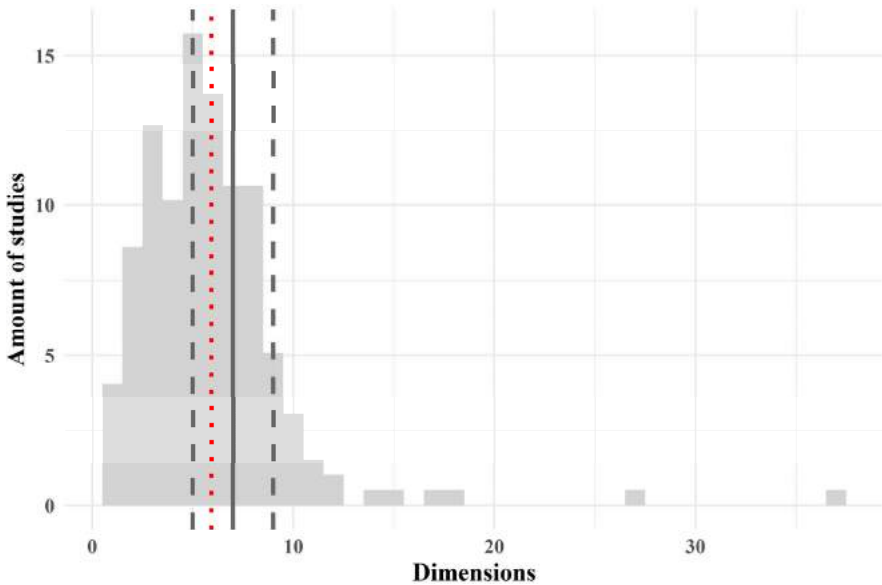
example, the findings of Auspurg and Jäckle (2017) imply that a large number of dimensions (e.g. 12 dimensions) can lead to order effects.¹⁰

As a consequence, researchers should avoid both too simple and too complex vignettes as well as unrealistic, implausible, and illogical scenarios (e.g. a professor without a school degree). Research shows that the use of such scenarios reduces the internal validity of inferences, because respondents no longer pay attention to the dimensions or, in the worst case, do not take the survey seriously (see Auspurg & Hinz, 2015a, pp. 40–42). That said, how many dimensions should approximately be provided to prevent boredom effects and cognitive overload among participants? The current state of research recommends seven dimensions to provide a good balance between simplicity and complexity (see Auspurg & Hinz, 2015a, pp. 18–22; Sauer et al., 2011). However, this is just a rough rule of thumb, since the choice should be guided by theory and depends on other factors such as research topic, survey length, respondents' motivation and cognitive skills as well as other FSE design aspects.

As Wallander (2009) reported, the number of dimensions in FSE studies published until 2006 has varied greatly between two (Steen & Cohen, 2004) and 25 (Thurman et al., 1988), with a median of six dimensions (see Wallander, 2009, p. 512). As Figure 6 indicates, this finding still holds for recent studies, which frequently deviate from this rough seven dimensions rule of thumb. While the average number of dimensions used in prior research since 2006 is 5.7, a substantial amount of research provides more than nine or less than five dimensions. Overall, 57% fall into the range of seven dimensions, while 38% of the publications provide fewer and 5% more vignette dimensions. Even after restricting the sample to FSE studies that are (a) more recently published and (b) have several ratings per person, we found that 42% of the studies use more or less vignette dimensions than suggested by the methodological literature.

Another important aspect concerning complexity is the *presentation style* of the vignette. Most researchers use text vignettes, while other forms of multifactorial survey experiments, such as conjoint analysis and choice experiments, often present FSE dimensions in tabular format. The cognitive load of reading a table is likely lower than reading a text with or without highlighted dimensions, which might affect response behaviour. Only recently has the first research about the differences between both presentation styles in FSE been published. Based on a student sample, Sauer et al. (2020) found no significant differences between presentation styles in relation to vignette rating and non-response. In contrast, Shamon et al. (2019) reported less non-response, in particular refusals, for a tabular presentation

10 Auspurg and Jäckle (2017) further found that respondents' degree of uncertainty about a topic influences the likelihood of order effects, while other studies discovered no (Robbins & Kiser, 2018)—or at least no strong—evidence for order effects of FSE dimensions (Düval & Hinz, 2020).



Note: The histogram shows the number of dimensions in FSE applications (2007–2018) with at least two ratings per person. The red dotted line displays the mean value of all included dimensions, a grey solid line displays the rule of thumb, and grey dashed lines denote the threshold of the rule of thumb.

Figure 6 Number of dimensions in FSE publications 2007-2018

than for textual scenario descriptions with and without underlining varied information, especially to the less-well educated people. Thus, given the results of these two studies that rely on different samples, the presentation style may affect response behaviour and data quality especially for less educated respondents but may matter less for other participants. Thus, keeping the limited number of studies in mind, one may cautiously conclude that using a tabular format may not hurt in some contexts, but may be beneficial depending on respondents' background. Since these are just first preliminary conclusions, more research needs to address under which circumstances—including the realism of the vignette and the complexity of the examined topic—the presentation style may affect the quality of the data.

Finally, the realism and complexity of a scenario also depend on the information provided and omitted. FSEs rely on the important yet underappreciated *assumption of information equivalence* (Dafoe et al., 2018). Participants need all relevant information necessary to assess a situation and to provide a meaningful answer. If a vignette lacks important aspects, respondents may update their beliefs and fill in the missing pieces in accordance with their expectations or stereotypes.

Given that respondents' expectations and stereotypes might not be exogenous to the individual background and the presented treatments, the lack of relevant information may lead to biased inferences about effects and causal mechanisms at work, which violates the assumption of information equivalence since individuals base their response on different information. For this reason, Dafoe et al. (2018, p. 406) proposed and evaluated three strategies for achieving information equivalence. The first strategy—encouraging respondents to think of an abstract instead of a real-world scenario—turned out to be ineffective. In contrast, the second strategy—using covariate control by specifying background details to prevent respondents from updating their beliefs—helped at least to reduce imbalance for the specified variables. The third strategy relies on framing the vignette scenario as the outcome of a random assignment process. Respondents are told that the treatment is the outcome of a random process (e.g. lottery, natural experiment) to make respondents believe that the treatment is not correlated with other, omitted dimensions, which may have an impact on the respondent's vignette rating. The third strategy turned out to be most effective in the study, while it remains open to future research to examine how effective this strategy is in other contexts. Irrespective of the findings of follow-up research, it becomes clear that design choices determine how realistic respondents perceive the described scenario to be, and how internally and externally valid inferences from the FSE will be. As our discussion underlines, this task is not just a technical exercise, but requires theoretical guidance and in-depth knowledge of the research topic under investigation.

Predictive Validity

Figure 7 depicts the number of articles in our analytical sample across time based on a stacked area chart. As Figure 7 shows, FSEs are increasingly used not only to study attitudes and hypothetical judgements (dark grey area), but also to explore the determinants of behavioural intentions (light grey area). On average, in each time period, approximately 45% of all FSE studies focus on behavioural intentions, with the strongest boost since 2010. For instance, FSEs are currently used to study willingness to pay (see Bekkers, 2010; Bridoux et al., 2016), hiring and job decisions (see Di Stasio & Gërkhani, 2015; van Belle et al., 2018), mobility behaviour (see Abraham et al., 2010; Teti et al., 2016), or medical and care decisions (see Drewniak et al., 2016; Shlay, 2010).

Obviously, an FSE does not measure actual behaviour, but asks participants to assess a hypothetical scenario based on the information provided. Hence, FSEs gauge self-reported behavioural intentions in a hypothetical situation. Thus, an important question regards the *predictive validity* of such measures (see Eifler & Petzold, 2019; Petzold & Wolbring, 2019): To what extent do hypothetical intentions correspond with real-world behaviour?

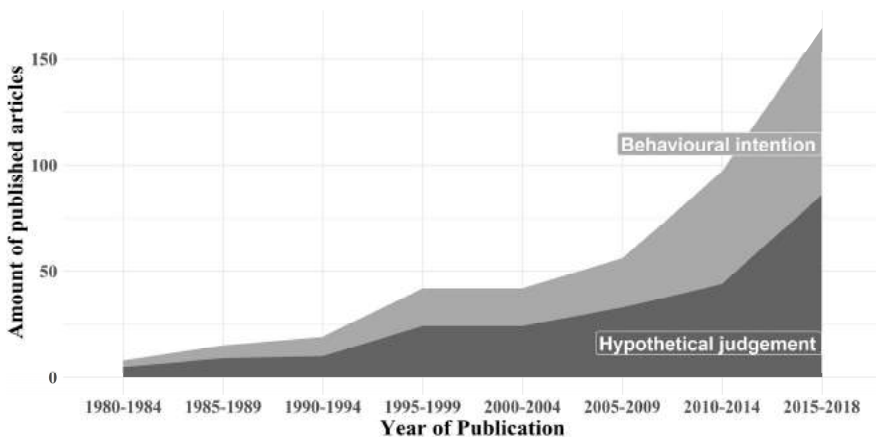


Figure 7 Behavioural intentions in FSE publications

A stated intention does not always correspond very well with real-world behaviour (see Barabas & Jerit, 2010; Collett & Childs, 2011). As the theory of planned behaviour (Fishbein & Ajzen, 2010) suggests, and as outlined by Petzold and Wolbring (2019) for FSEs, intentions may only translate into actual behaviour under certain conditions. For instance, actors might plan to act in a certain way, but in reality lack behavioural control or face the high costs of an action. For this reason, FSEs are sometimes criticised for lacking ‘*psychological realism*’ and predictive validity. In contrast, one could argue that although behavioural intentions do not perfectly predict real-world behaviour, they are important determinants of actual decision-making. Thus, showing what influences behavioural intentions might provide insights into the determinants of human action.

Unfortunately, the current state of research is small and inconclusive about the predictive validity of FSEs. Some evidence suggests low behavioural validity, with substantial differences between hypothetical decision-making and a behavioural benchmark regarding the distribution of the outcomes and their determinants (e.g. Pager & Quillian, 2005; Findley et al., 2017). In contrast, another group of studies concluded that FSEs have high predictive validity, and that the dimensions of an FSE sufficiently correspond with a behavioural benchmark (see Drasch, 2017; Hainmueller et al., 2015; Nisic & Auspurg, 2009; Raub & Buskens, 2008). Finally, a third group of studies offer results that are both partly in line with the first and the second position (e.g. Barabas & Jerit, 2010; Eifler, 2010; Petzold & Wolbring, 2019). They document that distributions of intended and actual behaviour clearly deviate from each other, indicating that other factors (such as social desirability and the costs of an action) co-determine decision-making in the real world. Despite the

reported differences in levels, these studies found that FSEs seem to provide correct estimates of behavioural determinants regarding direction and relative effect sizes (see, however, Barabas & Jerit, 2010).¹¹

Thus, the current state of research does not justify generally rejecting FSEs as a way of generating insights into determinants of behaviour, or using FSEs uncritically for all research questions in terms of behaviour. However, this ambiguous state of research raises several questions for future research that should be kept in mind when deciding whether to use an FSE to answer a specific research question and how to design it. In particular, the state of research raises the question: Under which conditions does an FSE have higher or lower predictive validity regarding human behaviour? Different factors must play a role in such theoretical considerations, including methodological aspects that affect the realism of the vignette, respondents' experience with the decision situation, and the sensitivity of the topic under investigation.

Concerning sensitivity, Wallander (2009) reported numerous FSEs on a wide range of topics that might be potentially affected by social desirability. Previous research has used FSEs to study *racial prejudice* (Shlay, 1986; St. John & Healdmoore, 1995), *sexual harassment* (Hunter & McClelland, 1991; Weber-Burdin & Rossi, 1982), and *drinking and driving* behaviour (Applegate et al., 1996; Thurman et al., 1993). Past research on sensitive questions and social desirability bias suggests that FSEs are better suited for studying sensitive questions than direct questions (see Alexander & Becker, 1978; Auspurg et al., 2015), and seem to outperform specific survey techniques such as the randomised response technique, which has been developed to attenuate social desirability bias in surveys (Armacost et al., 1991). Being better suited than other methods to attenuate social desirability bias does not imply that FSEs cannot suffer from social desirability bias and are adequate for examining sensitive topics. Social desirability might still be substantial and undermine causal inferences, especially if respondents become aware of the research topic. Respondents may quickly realise what the actual focus of the FSE is if the number dimensions is low, if the variation of the vignette is highlighted, or if several vignettes are rated sequentially in a within-subjects design. To our knowledge, only two studies have experimentally compared *within-subjects* with *between-subjects designs* with inconclusive results regarding the impact of

11 One important reason for this inconclusive state of research regarding the predictive validity FSEs might be that the reported validation studies rely on very different research designs, including within-person comparisons (e.g. Pager & Quillian, 2005), natural experiments (e.g. Hainmueller et al., 2015) and experimental designs (e.g. Petzold & Wolbring, 2019). Obviously, the limitations of a design for the estimation of a behavioural benchmark can result in biased estimates and undermine the validation strategy. Further, the measurement of the outcome, the sampling strategy of the FSE, and the behavioural benchmark differ in many of these validation studies, which might undermine comparability (see Petzold & Wolbring, 2019).

within-subjects designs in terms of social desirability (see Auspurg et al., 2015; Walzenbach, 2019). Given the small body of methodological research, we recommend conducting pre-tests to assess the sensitivity of a research topic or of an FSE dimension instead of relying on the general claim that an FSE is better equipped to address sensitive questions.

In addition, one might also suspect that FSEs have higher predictive power if a respondent is familiar with the described situation and if the scenario resembles the actual decision-making process in real life (Hainmueller et al., 2015). As a consequence, one might assume that moving decisions are well evaluated by respondents, who seriously consider or prepare an actual move. Planned behaviour may correspond well with actual behaviour in such an instance. In contrast, some survey participants might not even have in mind in which situation they perceive jaywalking to be acceptable. This may explain why past research has concluded that the same dimensions for the intention to move predict actual moving behaviour (e.g. Nisic & Auspurg, 2009), while predictive power is rather low in the case of jaywalking (Eifler, 2007). These considerations are speculative, but they illustrate that future research should focus more on the predictive validity of FSEs and the development of a theory specifying the conditions under which FSEs are informative about determinants of actual behaviour. To this end, more theory-driven validation studies appear promising, with systematic variation of the discussed factors.

Conclusions

This paper provides a literature review about the use of FSEs in the social sciences (1982–2018). Our literature review shows that the field of FSEs has developed rapidly since the mid-2000s. They are increasingly being applied in different research areas such as crime, care and health, work, and among scholars from different countries, in particular from the US, Germany, the Netherlands, and the UK. Approximately half of recent studies have relied on non-probability samples (such as convenience, referral, and purposive samples; and samples from the crowd-sourcing platform Amazon Mechanical Turk), raising questions about both the generalisability of results and the use of significance testing. Most recent studies have depended on within-subjects designs, and almost all have used state-of-the-art techniques to analyse such clustered data. In contrast, more recent advances in procedures for sampling vignette sets from a large vignette universe, such as *D-optimal* and *RBCF designs*, have not entered applied research to the same extent. While these techniques help to design FSEs in an order to avoid the confounding of main and interaction effects, and to optimise statistical power, they require additional expertise, specialised software, and time investment. Nonetheless, we especially recommend making extra investments in the case of small samples and

a large vignette universe, while the use of random sampling techniques still leads to inefficiencies and untestable assumptions, but might be acceptable in the case of very large sample from the vignette universe.

Several methodological questions remain unresolved concerning the realism and complexity of vignettes, social desirability, and the predictive validity of FSEs with respect to human behaviour. Regarding the *complexity and realism of vignettes*, we focused on the number of dimensions in an FSE and highlighted that simple scenarios may lead to boredom and fatigue effects, while very detailed scenarios may seem realistic but cause cognitive overload among respondents. However, a 'one-size-fits-all' rule regarding the complexity of vignettes and the number of vignette dimensions does not exist and is unlikely to emerge in the future. Thus, the design of factorial surveys should rely on theoretical considerations, and not just on considerations related to technical aspects of the experimental design. For example, researchers need to take into account the individual background, motivation and cognitive skills of their respondents as well as peculiarities of their research topic. Furthermore, the complexity of a vignette design and the related cognitive load depend not only on the number of dimensions (ratings), but also on other design elements, such as the measurement of the outcome and the vignette presentation style (e.g. Sauer et al., 2020).

In a similar vein, some researchers have used video vignettes to present scenarios. Audio-visual stimuli seem very promising and have the potential to increase the realism of vignettes substantially. Nevertheless, researchers should be aware that conducting video vignettes is demanding and may introduce new methodological pitfalls, such as the confounding of vignette dimensions with the (non)verbal expressions of the actors. Video vignettes may also be prone to other well-known methodological aspects (such as social desirability) due to the salience of certain vignette dimensions (see Ceuterick et al., 2020). Hence, researchers should carefully consider which presentation style seems most adequate. Instead of a 'one-size-fits-all' rule, we recommend relying on theoretical considerations and pre-tests to assess the various FSE design aspects. For example, theory can help to identify potential interactions between the research topic, the number of ratings per person, and the number of dimensions. Moreover, participants might be differently affected depending upon their motivation to take part in the survey, their cognitive skills, previous experience, and familiarity with the described situation (see Sauer et al., 2011; Teti et al., 2016).

Further, our review shows that FSEs are not only increasingly being used to study attitudes, but also to explore the determinants of behaviour, and in each observed time period, approximately 45% of all FSE studies focus on behaviour as an outcome. We indicated that the current state of research is inconclusive and raises several methodological and theoretical challenges for future validation studies. These questions illustrate that future research should aim to integrate previous

research and to formulate a theory that specifies the conditions under which FSEs are informative about the determinants of actual behaviour. As long as such theory does not exist, it appears neither warranted to reject FSEs to generate insights into determinants of behaviour, nor to apply them uncritically. When making inferences from stated intentions in FSEs to actual behaviour in the real world, potential differences need to be considered, such as the possibility of an intention-behaviour gap, respondents' lack of familiarity with the decision situation, and biases due to social desirability.

Finally, there is a need for better documentation and reporting standards to assess the methodological aspects of FSEs. In a substantial number of publications, key information about the FSE design was hard to find, buried in online appendices, or not reported at all. In particular, it was alarming that an increasing number of recent publications did not contain information about how the vignettes were sampled from the universe. The fact that we could not retrieve this information, even after an extensive search, is alarming and indicates the need to establish clear documentation and reporting standards for FSEs. This includes all methodological aspects that are necessary to assess the quality of an instrument and to conduct replications and follow-up studies.

References

- Abraham, M., Auspurg, K., & Hinz, T. (2010). Migration Decisions within Dual-earner Partnerships: A Test of Bargaining Theory. *Journal of Marriage and Family*, 72(4), 876–892. <https://doi.org/10.1111/j.1741-3737.2010.00736.x>
- Adamle, K. N., Ludwick, R., Zeller, R., & Winchell, J. (2008). Oncology Nurses' Responses to Patient-initiated Humor. *Cancer Nursing*, 31(6), E1-9. <https://doi.org/10.1097/01.NCC.0000339243.51291.cc>
- Alexander, C. S., & Becker, H. J. (1978). The Use of Vignettes in Survey Research. *Public Opinion Quarterly*, 42(1), 93–104. <https://doi.org/10.1086/268432>
- Applegate, B. K., Cullen, F. T., Link, B. G., Richards, P. J., & Lanza-Kaduce, L. (1996). Determinants of Public Punitiveness toward Drunk Driving: A Factorial Survey Approach. *Justice Quarterly*, 13(1), 57–79. <https://doi.org/10.1080/07418829600092821>
- Armocost, R. L., Hosseini, J. C., Morris, S. A., & Rehbein, K. A. (1991). An Empirical Comparison of Direct Questioning, Scenario, and Randomized Response Methods for Obtaining Sensitive Business Information. *Decision Sciences*, 22(5), 1073–1090. <https://doi.org/10.1111/j.1540-5915.1991.tb01907.x>
- Atzmüller, C., & Steiner, P. M. (2010). Experimental Vignette Studies in Survey Research. *Methodology*, 6(3), 128–138. <https://doi.org/10.1027/1614-2241/a000014>
- Auspurg, K., & Hinz, T. (2015a). *Factorial Survey Experiments*. Sage.
- Auspurg, K., & Hinz, T. (2015b). Multifactorial experiments in surveys: Conjoint analysis, choice experiments, and factorial surveys. In M. Keuschnigg & T. Wolbring (Eds.), *Soziale Welt Sonderband: Vol. 22. Experimente in den Sozialwissenschaften* (1st ed., pp. 291–315). Nomos.

- Auspurg, K., Hinz, T., & Sauer, C. (2017). Why Should Women Get Less? Evidence on the Gender Pay Gap from Multifactorial Survey Experiments. *American Sociological Review*, 82(1), 179–210. <https://doi.org/10.1177/0003122416683393>
- Auspurg, K., Hinz, T., Sauer, C., & Liebig, S. (2015). The factorial survey as method for measuring sensitive issues. In U. Engel, B. Jann, P. Lynn, A. C. Scherpenzeel, & P. Sturgis (Eds.), *Improving Survey Methods: Lessons from Recent Research* (pp. 137–149). Routledge Taylor & Francis Group.
- Auspurg, K., & Jäckle, A. (2017). First Equals Most Important? Order Effects in Vignette-Based Measurement. *Sociological Methods & Research*, 46(3), 490–539. <https://doi.org/10.1177/0049124115591016>
- Aviram, H. (2012). What would you do? Conducting web-based factorial vignette surveys. In L. Gideon (Ed.), *Handbook of Survey Methodology for the Social Sciences* (pp. 463–473). Springer.
- Bader, F., Baumeister, B., Berger, R., & Keuschnigg, M. (2019). On the Transportability of Laboratory Results. *Sociological Methods & Research*, 62. <https://doi.org/10.1177/0049124119826151>
- Baker, P. M. (1983). Ageism, Sex, and Age: A Factorial Survey Approach. *Canadian Journal on Aging*, 2(4), 177–184. <https://doi.org/10.1017/S0714980800004645>
- Barabas, J., & Jerit, J. (2010). Are Survey Experiments Externally Valid? *American Political Science Review*, 104(2), 226–242. <https://doi.org/10.1017/S0003055410000092>
- Baron, S. W., Forde, D. R., & Kennedy, L. W. (2001). Rough Justice: Street Youth and Violence. *Journal of Interpersonal Violence*, 16(7), 662–678. <https://doi.org/10.1177/088626001016007003>
- Baughman, K. R., Ludwick, R., Jarjoura, D., Kropp, D., & Shenoy, V. (2019). Advance Care Planning in Skilled Nursing Facilities: A Multisite Examination of Professional Judgments. *The Gerontologist*, 59(2), 338–346. <https://doi.org/10.1093/geront/gnx129>
- Bekkers, R. (2010). Who gives what and when? A Scenario Study of Intentions to Give Time and Money. *Social Science Research*, 39(3), 369–381. <https://doi.org/10.1016/j.ssresearch.2009.08.008>
- Bell, M. L., & Forde, D. R. (1999). A Factorial Survey of Interpersonal Conflict Resolution. *The Journal of Social Psychology*, 139(3), 369–377. <https://doi.org/10.1080/00224549909598392>
- Brenner, M., O’Shea, M., J Larkin, P., Kamionka, S. L., Berry, J., Hiscock, H., Rigby, M., & Blair, M. (2017). Exploring Integration of Care for Children Living with Complex Care Needs across the European Union and European Economic Area. *International Journal of Integrated Care*, 17(2), 1. <https://doi.org/10.5334/ijic.2544>
- Bridoux, F., Stofberg, N., & Den Hartog, D. (2016). Stakeholders’ Responses to CSR Tradeoffs: When Other-Oriented and Trust Trump Material Self-Interest. *Frontiers in Psychology*, 6(1992), 1–18. <https://doi.org/10.3389/fpsyg.2015.01992>
- Buskens, V., & Weesie, J. (2000). An Experiment on the Effects of Embeddedness in Trust Situations: Buying a Used Car. *Rationality and Society*, 12(2), 227–253. <https://doi.org/10.1177/104346300012002004>
- Cahan, S. F. (1996). Political Use of Income: Some Experimental Evidence from Capitol Hill. *The Journal of Socio-Economics*, 25(1), 69–87. [https://doi.org/10.1016/S1053-5357\(96\)90054-2](https://doi.org/10.1016/S1053-5357(96)90054-2)
- Ceuterick, M., Bracke, P., van Canegem, T., & Buffel, V. (2020). Assessing Provider Bias in General Practitioners’ Assessment and Referral of Depressive Patients with Different

- Migration Backgrounds: Methodological Insights on the Use of a Video-Vignette Study. *Community Mental Health Journal*, 56(8), 1457–1472.
<https://doi.org/10.1007/s10597-020-00590-y>
- Chatfield, S. L., Gamble, A., & Hallam, J. S. (2018). Men's Preferences for Physical Activity Interventions: An Exploratory Study Using a Factorial Survey Design Created With R Software. *American Journal of Men's Health*, 12(2), 347–358.
<https://doi.org/10.1177/1557988316643316>
- Collett, J. L., & Childs, E. (2011). Minding the Gap: Meaning, Affect, and the Potential Shortcomings of Vignettes. *Social Science Research*, 40(2), 513–522.
<https://doi.org/10.1016/j.ssresearch.2010.08.008>
- Couper, M. P., & Singer, E. (2012). Informed Consent for Web Paradata Use. *Survey Research Methods*, 7(1), 57–67. <https://doi.org/10.18148/srm/2013.v7i1.5138>
- Dafoe, A., Zhang, B., & Caughey, D. (2018). Information Equivalence in Survey Experiments. *Political Analysis*, 26(4), 399–416. <https://doi.org/10.1017/pan.2018.9>
- Deaton, A., & Cartwright, N. (2018). Understanding and Misunderstanding Randomized Controlled Trials. *Social Science & Medicine*, 210, 2–21.
<https://doi.org/10.1016/j.socscimed.2017.12.005>
- Di Stasio, V., & Gërxxhani, K. (2015). Employers' Social Contacts and their Hiring Behavior in a Factorial Survey. *Social Science Research*, 51(1), 93–107.
<https://doi.org/10.1016/j.ssresearch.2014.12.015>
- Drasch, K. (2017). Behavioral Intentions, Actual Behavior and the Role of Personality Traits: Evidence from a Factorial Survey among Female Labor Market Re-Entrants. *Methods, Data, Analysis*, 13(2), 1–23. <https://doi.org/10.12758/mda.2017.14>
- Drewniak, D., Krones, T., Sauer, C., & Wild, V. (2016). The Influence of Patients' Immigration Background and Residence Permit Status on Treatment Decisions in Health Care: Results of a Factorial Survey among General Practitioners in Switzerland. *Social Science & Medicine*, 161, 64–73. <https://doi.org/10.1016/j.socscimed.2016.05.039>
- Dülmer, H. (2007). Experimental Plans in Factorial Surveys: Random or Quota Design? *Sociological Methods & Research*, 35(3), 382–409.
<https://doi.org/10.1177/0049124106292367>
- Dülmer, H. (2016). The Factorial Survey: Design Selection and its Impact on Reliability and Internal Validity. *Sociological Methods & Research*, 45(2), 304–347. <https://doi.org/10.1177/0049124115582269>
- Düval, S., & Hinz, T. (2020). Different Order, Different Results? The Effects of Dimension Order in Factorial Survey Experiments. *Field Methods*, 32(1), 23–37.
<https://doi.org/10.1177/1525822X19886827>
- Eifler, S. (2007). Evaluating the Validity of Self-Reported Deviant Behavior Using Vignette Analyses. *Quality & Quantity*, 41(2), 303–318.
<https://doi.org/10.1007/s11135-007-9093-3>
- Eifler, S. (2010). Validity of a Factorial Survey Approach to the Analysis of Criminal Behavior. *Methodology*, 6(3), 139–146. <https://doi.org/10.1027/1614-2241/a000015>
- Eifler, S., & Petzold, K. (2019). Validity aspects of vignette experiments: Expected “what-if” differences between reports of behavioral intentions and actual behavior. In P. J. Lavrakas, M. W. Traugott, C. Kennedy, A. L. Holbrook, E. D. de Leeuw, & Brady T. West (Eds.), *Experimental Methods in Survey Research: Techniques that Combine Random Sampling with Random Assignment* (pp. 393–416). John Wiley & Sons, Ltd.
<https://doi.org/10.1002/9781119083771.ch20>

- Findley, M. G., Laney, B., Nielson, D. L., & Sharman, J. C. (2017). External Validity in Parallel Global Field and Survey Experiments on Anonymous Incorporation. *The Journal of Politics*, 79(3), 856–872. <https://doi.org/10.1086/690615>
- Fishbein, M., & Ajzen, I. (2010). *Predicting and Changing Behavior: The Reasoned Action Approach*. Psychology Press.
- Graeff, P., Sattler, S., Mehlkop, G., & Sauer, C. (2014). Incentives and Inhibitors of Abusing Academic Positions: Analysing University Students' Decisions about Bribing Academic Staff. *European Sociological Review*, 30(2), 230–241. <https://doi.org/10.1093/esr/jct036>
- Haase, M., Becker, I., Nill, A., Shultz, C. J., & Gentry, J. W. (2016). Male Breadwinner Ideology and the Inclination to Establish Market Relationships: Model Development Using Data from Germany and a Mixed-Methods Research Strategy. *Journal of Macromarketing*, 36(2), 149–167. <https://doi.org/10.1177/0276146715576202>
- Hainmueller, J., Hangartner, D., & Yamamoto, T. (2015). Validating Vignette and Conjoint Survey Experiments against Real-world Behavior. *Proceedings of the National Academy of Sciences*, 112(8), 2395–2400. <https://doi.org/10.1073/pnas.1416587112>
- Hennessy, M., MacQueen, K. M., & Seals, B. (1995). Using Factorial Surveys for Designing Intervention Programs. *Evaluation Review*, 19(3), 294–312. <https://doi.org/10.1177/0193841X9501900304>
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, 466(7302), 29. <https://doi.org/10.1038/466029a>
- Hunter, C., & McClelland, K. (1991). Honoring Accounts for Sexual Harassment: A Factorial Survey Analysis. *Sex Roles*, 24(11-12), 725–752. <https://doi.org/10.1007/BF00288209>
- Imbens, G., & Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press. <https://doi.org/10.1017/CBO9781139025751>
- Jackson, M., & Cox, D. R. (2013). The Principles of Experimental Design and Their Application in Sociology. *Annual Review of Sociology*, 39(1), 27–49. <https://doi.org/10.1146/annurev-soc-071811-145443>
- Jasso, G. (2006). Factorial Survey Methods for Studying Beliefs and Judgments. *Sociological Methods & Research*, 34(3), 334–423. <https://doi.org/10.1177/0049124105283121>
- Jasso, G., & Rossi, P. H. (1977). Distributive Justice and Earned Income. *American Sociological Review*, 42(4), 639. <https://doi.org/10.2307/2094561>
- Jörg, F., Borgers, N., Schrijvers, A. J. P., & Hox, J. J. (2006). Variation in Long-term Care Needs Assessors' Willingness to Support Clients' Requests for Admission to a Residential Home: A Vignette Study. *Journal of Aging and Health*, 18(6), 767–790. <https://doi.org/10.1177/0898264306293605>
- Kessler, T. M., Maric, A., Mordasini, L., Wöllner, J., Pannek, J., Mehnert, U., van Kerrebroeck, P. E., & Bachmann, L. M. (2014). Urologists' Referral Attitude for Sacral Neuromodulation for Treating Refractory Idiopathic Overactive Bladder Syndrome: Discrete Choice Experiment. *Neurourology and Urodynamics*, 33(8), 1240–1246. <https://doi.org/10.1002/nau.22490>
- Kiesewetter, I., Könings, K. D., Kager, M., & Kiesewetter, J. (2018). Undergraduate Medical Students' Behavioural Intentions towards Medical Errors and How to Handle Them: A Qualitative Vignette Study. *BMJ Open*, 8(3). <https://doi.org/10.1136/bmjopen-2017-019500>

- Krosnick, J. A. (1991). Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied Cognitive Psychology*, 5(3), 213–236. <https://doi.org/10.1002/acp.2350050305>
- Kuhfeld, W. F., Tobias, R. D., & Garratt, M. (1994). Efficient Experimental Design with Marketing Research Applications. *Journal of Marketing Research*, 31(4), 545–557. <https://doi.org/10.2307/3151882>
- Lancaster, K. J. (1966). A New Approach to Consumer Theory. *Journal of Political Economy*, 74(2), 132–157. <https://doi.org/10.1086/259131>
- Lepièce, B., Dubois, T., Jacques, D., & Zdanowicz, N. (2018). “Please admire me!” When Healthcare Providers’ Positive Stereotypes of Asylum Seeker Patients Contribute to Better Continuity of Care. *Psychiatria Danubina*, 30(7), 498–501.
- Liebe, U., & Meyerhoff, J. (2021). Mapping Potentials and Challenges of Choice Modelling for Social Science Research. *Journal of Choice Modelling*, 38 (Special Issue on Choice Modelling in Social Science Research), 100270. <https://doi.org/10.1016/j.jocm.2021.100270>
- Liebig, S., Sauer, C., & Friedhoff, S. (2015). Using Factorial Surveys to Study Justice Perceptions: Five Methodological Problems of Attitudinal Justice Research. *Social Justice Research*, 28(4), 415–434. <https://doi.org/10.1007/s11211-015-0256-4>
- Love, M. B., Davoli, G. W., & Thurman, Q. C [Q. C.] (1996). Normative Beliefs of Health Behavior Professionals regarding the Psychosocial and Environmental Factors That Influence Health Behavior Change Related to Smoking Cessation, Regular Exercise, and Weight Loss. *American Journal of Health Promotion*, 10(5), 371–379. <https://doi.org/10.4278/0890-1171-10.5.371>
- Ludwick, R., Wright, M. E., Zeller, R. A., Dowding, D. W., Lauder, W., & Winchell, J. (2004). An Improved Methodology for Advancing Nursing Research: Factorial Surveys. *Advances in Nursing Science*, 27(3), 224–238. <https://doi.org/10.1097/00012272-200407000-00007>
- Lyons, C. J. (2008). Individual Perceptions and the Social Construction of Hate Crimes: A Factorial Survey. *The Social Science Journal*, 45(1), 107–131. <https://doi.org/10.1016/j.soscij.2007.12.013>
- Maas, C. J., & Hox, J. J. (2004). The Influence of Violations of Assumptions on Multilevel Parameter Estimates and their Standard Errors. *Computational Statistics & Data Analysis*, 46(3), 427–440. <https://doi.org/10.1016/j.csda.2003.08.006>
- Manski, C. F. (1977). The Structure of Random Utility Models. *Theory and Decision*, 8(3), 229–254. <https://doi.org/10.1007/BF00133443>
- McFadden, D. (1974). Conditional Logit Analysis of Qualitative Choice Behavior. In *Frontiers in Econometrics*, Hrsg. Paul Zarembka, 105–142. New York: Academic.
- Moynihan, D. P. (2013). Does Public Service Motivation Lead to Budget Maximization? Evidence from an Experiment. *International Public Management Journal*, 16(2), 179–196. <https://doi.org/10.1080/10967494.2013.817236>
- Mutz, D. C. (2011). *Population-based Survey Experiments*. Princeton University Press. <https://doi.org/10.1515/9781400840489>
- Nisic, N., & Auspurg, K. (2009). Faktorieller Survey und klassische Bevölkerungsumfrage im Vergleich: Validität, Grenzen und Möglichkeiten beider Ansätze. In P. Kriwy & C. Voss (Eds.), *Klein aber fein! Quantitative empirische Sozialforschung mit kleinen Fallzahlen* (pp. 211–245). VS Verlag für Sozialwissenschaften.
- Oberoi, D. V., Jiwa, M., McManus, A., & Parsons, R. (2016). Do Men Know Which Lower Bowel Symptoms Warrant Medical Attention? A Web-based Video Vignette Survey of

- Men in Western Australia. *American Journal of Men's Health*, 10(6), 474–486.
<https://doi.org/10.1177/1557988315574739>
- Opp, K.D. (2002). When Do Norms Emerge by Human Design and When by the Unintended Consequences of Human Action? The Example of the No-smoking Norm. *Rationality and Society*, 14(2), 131–158. <https://doi.org/10.1177/1043463102014002001>
- Pager, D., & Quillian, L. (2005). Walking the Talk? What Employers Say Versus What They Do. *American Sociological Review*, 70(3), 355–380.
<https://doi.org/10.1177/000312240507000301>
- Peters, P., & Dulk, L. den (2003). Cross Cultural Differences in Managers' Support for Home-Based Telework. *International Journal of Cross Cultural Management*, 3(3), 329–346. <https://doi.org/10.1177/1470595803003003005>
- Petzold, K., & Wolbring, T. (2019). What Can We Learn From Factorial Surveys About Human Behavior? *Methodology*, 15(1), 19–30. <https://doi.org/10.1027/1614-2241/a000161>
- Raub, W., & Buskens, V. (2008). Theory and Empirical Research in Analytical Sociology: The Case of Cooperation in Problematic Social Situations. *Analyse & Kritik*, 30(2), 453. <https://doi.org/10.1515/auk-2008-0218>
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage.
- Reisel, A. (2017). Practitioners' Perceptions and Decision-making Regarding Child Sexual Exploitation - A Qualitative Vignette Study. *Child & Family Social Work*, 22(3), 1292–1301. <https://doi.org/10.1111/cfs.12346>
- Rix, J., Sheehy, K., Fletcher-Campbell, F., Crisp, M., & Harper, A. (2013). Exploring Provision for Children Identified with Special Educational Needs: An International Review of Policy and Practice. *European Journal of Special Needs Education*, 28(4), 375–391. <https://doi.org/10.1080/08856257.2013.812403>
- Robbins, B. G., & Kiser, E. (2018). Legitimate Authorities and Rational Taxpayers: An Investigation of Voluntary Compliance and Method Effects in a Survey Experiment of Income Tax Evasion. *Rationality and Society*, 30(2), 247–301.
<https://doi.org/10.1177/1043463118759671>
- Rosenbaum, P. R. (2010). *Design of Observational Studies*. Springer.
<https://doi.org/10.1007/978-1-4419-1213-8>
- Rossi, P. H., Sampson, W. A., Bose, C. E., Jasso, G., & Passel, J. (1974). Measuring Household Social Standing. *Social Science Research*, 3(3), 169–190.
[https://doi.org/10.1016/0049-089X\(74\)90011-8](https://doi.org/10.1016/0049-089X(74)90011-8)
- Sampson, W. A., & Rossi, P. H. (1975). Race and Family Social Standing. *American Sociological Review*, 40(2), 201. <https://doi.org/10.2307/2094345>
- Sauer, C., Auspurg, K., & Hinz, T. (2020). Designing Multi-Factorial Survey Experiments: Effects of Presentation Style (Text or Table), Answering Scales, and Vignette Order. *Methods, Data, Analysis*, 14(2), 195–214. <https://doi.org/10.12758/MDA.2020.06>
- Sauer, C., Auspurg, K., Hinz, T., & Liebig, S. (2011). The Application of Factorial Surveys in General Population Samples: The Effects of Respondent Age and Education on Response Times and Response Consistency. *Survey Research Methods*, 5(3), 89–102. <https://doi.org/10.18148/srm/2011.v5i3.4625>
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Houghton Mifflin.
- Shamon, H., Dülmer, H., & Giza, A. (2019). The Factorial Survey: The Impact of the Presentation Format of Vignettes on Answer Behavior and Processing Time. *Sociological*

- Methods & Research, First published online (June 25, 2019).*
<https://doi.org/10.1177/0049124119852382>
- Shlay, A. (1986). Taking Apart the American Dream: The Influence of Income and Family Composition on Residential Evaluations. *Urban Studies*, 23(4), 253–270.
<https://doi.org/10.1080/00420988620080331>
- Shlay, A. (2010). African American, White and Hispanic child care preferences: A Factorial Survey Analysis of Welfare Leavers by Race and Ethnicity. *Social Science Research*, 39(1), 125–141. <https://doi.org/10.1016/j.ssresearch.2009.07.005>
- Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Sage.
- St. John, C., & Healdmoore, T. (1995). Fear of Black Strangers. *Social Science Research*, 24(3), 262–280. <https://doi.org/10.1006/ssre.1995.1010>
- Steen, S., & Cohen, M. A. (2004). Assessing the Public's Demand for Hate Crime Penalties. *Justice Quarterly*, 21(1), 91–124. <https://doi.org/10.1080/07418820400095751>
- Su, D., & Steiner, P. M. (2020). An Evaluation of Experimental Designs for Constructing Vignette Sets in Factorial Surveys. *Sociological Methods & Research*, 49(2), 1–43.
<https://doi.org/10.1177/0049124117746427>
- Taylor, B. J. (2005). Factorial Surveys: Using Vignettes to Study Professional Judgement. *British Journal of Social Work*, 36(7), 1187–1207. <https://doi.org/10.1093/bjsw/bch345>
- Teele, D. L. (2014). *Field Experiments and their Critics: Essays on the Uses and Abuses of Experimentation in the Social Sciences*. Yale University Press.
- Teti, A., Gross, C., Knoll, N., & Blüher, S. (2016). Feasibility of the Factorial Survey Method in Aging Research: Consistency Effects among Older Respondents. *Research on Aging*, 38(7), 715–741. <https://doi.org/10.1177/0164027515600767>
- Thurman, Q. C [Quint C.], Jackson, S., & Zhao, J. (1993). Drunk-Driving Research and Innovation: A Factorial Survey Study of Decisions to Drink and Drive. *Social Science Research*, 22(3), 245–264. <https://doi.org/10.1006/ssre.1993.1012>
- Thurman, Q. C [Quint C.], Lam, J. A., & Rossi, P. H. (1988). Sorting Out the Cuckoo's Nest: A Factorial Survey Approach to the Study of Popular Conceptions of Mental Illness. *The Sociological Quarterly*, 29(4), 565–588.
<https://doi.org/10.1111/j.1533-8525.1988.tb01435.x>
- Tolsma, J., Blaauw, J., & te Grotenhuis, M. (2012). When Do People Report Crime to the Police? Results from a Factorial Survey Design in the Netherlands, 2010. *Journal of Experimental Criminology*, 8(2), 117–134. <https://doi.org/10.1007/s11292-011-9138-4>
- van Belle, E., Di Stasio, V., Caers, R., Couck, M. de, & Baert, S. (2018). Why Are Employers Put Off by Long Spells of Unemployment? *European Sociological Review*, 34(6), 694–710. <https://doi.org/10.1093/esr/jcy039>
- van der Sluis, M. E., Reezigt, G. J., & Borghans, L. (2014). Quantifying Stakeholder Values of VET Provision in the Netherlands. *Vocations and Learning*, 7(1), 1–19.
<https://doi.org/10.1007/s12186-013-9104-6>
- Wallander, L. (2009). 25 Years of Factorial Surveys in Sociology: A Review. *Social Science Research*, 38(3), 505–520. <https://doi.org/10.1016/j.ssresearch.2009.03.004>
- Walzenbach, S. (2019). Hiding Sensitive Topics by Design? An Experiment on the Reduction of Social Desirability Bias in Factorial Surveys. *Survey Research Methods*, 13(1), 103–121. <https://doi.org/10.18148/SRM/2019.V1I1.7243>

- Weber-Burdin, E., & Rossi, P. H. (1982). Defining Sexual Harassment on Campus: A Replication and Extension. *Journal of Social Issues*, 38(4), 111–120. <https://doi.org/10.1111/j.1540-4560.1982.tb01913.x>
- Wouters, R., & Walgrave, S. (2017). Demonstrating Power. *American Sociological Review*, 82(2), 361–383. <https://doi.org/10.1177/0003122417690325>

Controlling for Taste Preferences – A Factorial Survey about the Orientation to Judgment Devices in Movie Choice

Clemens Maria Schmidt

University of Trier

Abstract

This paper examines the gains in complexity reduction and causality identification provided by the factorial survey for the analysis of a market characterized by uncertainty. The starting point is the problem of quality uncertainty for the market actor, commonly dealt with in economic sociology. Using Karpik's approach of the 'Economics of Singularities', the problem of choosing the right movie is expounded and the question of what moviegoers base their choice on is developed. The uncertainty in question is the result of subjective tastes, which also leads to a methodological problem. As a result, previous studies measured taste preferences instead of the influence of judgment devices. By means of a study on the right choice of movie, the paper shows that the method of the factorial survey has the important advantages of being able to control for taste preferences as well as to detect causality. Data collected among students is presented and hypotheses based on Karpik's concepts are tested. The results show that expert judgements such as critics' recommendations and awards have a high influence on the choice of independent movies. On the other hand, the choice of blockbuster movies is additionally influenced by its listing in the charts and the ratings by other consumers. This shows not only that different social devices are used for orientation depending on preference, but also how strong their influence is in each case. Therefore, it is argued that the factorial survey method offers some advantages for the analysis of the causal influence of judgment devices in choice situations, especially for singular goods, which are highly complex and thus difficult to compare. Finally, limitations of the study as well as the method used are discussed.

Keywords: Economic Sociology, Factorial Survey, Judgment Devices, Markets, Movie Market, Taste, Uncertainty



© The Author(s) 2021. This is an Open Access article distributed under the terms of the Creative Commons Attribution 3.0 License. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

A core concern of economic sociology is to show that economic actors are confronted with uncertainty and to ask how the actors (can) deal with it and what phenomena result from it (Beckert, 1996; Granovetter, 1985; Maurer & Schmidt, 2019). A wide range of markets are analyzed in an extensive literature. Both for supposedly strictly rational financial markets (Preda, 2007) and for markets on which cultural products are traded (Aspers, 2016; Dekker & de Jong, 2016; Keuschnigg, 2015) it is shown that suppliers and buyers are confronted with uncertainty. Especially when the quality of a good or service is unclear before purchase—in economics one speaks of experience goods (P. Nelson, 1970)—it should be clarified how consumers make a decision for a certain product and which market structure emerges. Various studies show that the problem of decision making is solved with the help of different social mechanisms. For example, the social network is used to obtain trustworthy information (Keuschnigg, 2015) or the status of the supplier is used to select a high-quality product (Benjamin & Podolny, 1999; Podolny, 2008). Several empirical methods are used for this purpose: Network analyses, qualitative interviews, and statistical analyses of macro data and standardized questionnaires. Explanatory experimental designs, on the other hand, are hardly ever applied to questions of action choice under quality uncertainty in markets (Kittel, 2015).

Lucien Karpik (2010) presents a theoretical approach that has been very positively received by the scientific community¹. He shows that the quality problem is particularly relevant for cultural products, but also for services provided by e.g., doctors or lawyers, which are difficult to evaluate before purchase because these products are unique. He argues that in such cases, depending on the market, certain judgment devices offer orientation, on the basis of which he identifies different regimes. Some markets, for example, are based on the purchasing decisions of many other consumers, e.g., music charts in the common-opinion regime, while in other markets, e.g., literature, expert judgments are more important (the expert-opinion regime). Despite the positive discussion among scholars, little further work has been done based on Karpik's considerations and few empirical studies

1 The French original was published in 2007, the English edition in 2010 and the German one in 2011. For the academic discussion in French-speaking countries see Gadrey (2008), for the English Campbell (2010); Espeland (2011); Healy (2011), Hutter (2011a, 2011b) and for the German Kraemer (2017); Maurer (2014).

Acknowledgements

The author would like to thank the two anonymous reviewers whose comments helped improve and clarify this manuscript, and Christopher Dorn whose annotations made the text more reader-friendly.

Direct correspondence to

Clemens Maria Schmidt, University of Trier, Germany
E-mail: schmidtc@uni-trier.de

are quantitative². As will be shown in the following, the reason for this lies in the problem of treating such complex phenomena empirically. Karpik himself foresaw this difficulty when he wrote “the main obstacle is perhaps less theoretical than methodological and technical” and concluded with the question: “How are we to observe and identify the effects of configurations of competing impersonal judgment devices whose scales and means of action are highly diverse?” (Karpik, 2010, p. 256). This paper thus addresses two related problems. On the one hand, the economic sociological problem of choice under uncertainty with the ensuing question of what market structure exists. A major cause of this problem is the relevance of subjective taste, which makes it difficult to determine the quality *ex ante* of an action in a complex situation. On the other hand, there is the methodological problem of how to empirically measure the causal effect of influencing variables without actually measuring taste preferences. Here the proposal is made and discussed to what extent this question can be answered with the factorial survey. For this purpose, the problem of quality uncertainty in the movie market is addressed: How to choose the right cinema movie?

The general idea underlying this question here is that a choice of action depends on desires, beliefs, and opportunities (Hedström, 2005). The goal of seeing a movie in the cinema that one likes can be secondary to the desire to do something with friends. Also, the possible action alternatives are to be seen in relation to the concrete desires. For example, if the primary goal is to be entertained, options such as going to the theater or a concert are also conceivable. The question pursued here about the choice of a movie can therefore be seen in different settings. One case would be that the overriding goal is to go to the cinema (rather than the soccer stadium) and now there are several movies to choose from. Another, more general perspective does not necessarily see this decision to go to the cinema as having already been made, but supposes that a movie gains attention and the decision is then made whether to watch it in the cinema—with whom and when is downstream. In one case, one is already standing at the ticket counter and must only decide on one of the movies on offer, and in the other case, the desire to go to the cinema must first be elicited. Here, the question is pursued with the aim of finding out what drives the belief that going to the cinema for a movie is the right choice.

In the next section, the state of research on the choice of a movie is summarized and, based on the considerations of Lucien Karpik, the problem is discussed, its solution model is presented and hypotheses are developed. The design of the factorial survey for the empirical testing of the problem is presented in the following section and the results are then presented. The last section concludes with a

2 Among the works that use Karpik's approach is the case study of film evaluations by Bialecki et al. (2017). For a Karpik-based and quantitative analysis of the wine market, see Schenk (2021).

discussion of the advantages and disadvantages of the factorial survey for questions of choice under quality uncertainty.

The Movie Market and Karpik's Economics of Singularities: Theory and Hypotheses

The Problem of Uncertainty in the Movie Market: Nobody knows anything

Movies are not only a medium of entertainment, a venue for social debates, and a political instrument for forming opinions; as a product of the creative industries, movies are also economically relevant. Driven by technical progress and guided by social processes, there is now a high-turnover market for the production, distribution, and exhibition of movies (Scott, 2005). On these three levels, two types can be distinguished structurally. The market is divided into a larger blockbuster market with financially strong corporations and a smaller art house and independent market in which less popular niches are served (de Valck, 2007, pp. 128-130). However, they are united by the problem of uncertainty that structures the entire market. For whether a movie produced for a lot of money will be successful (and pay off financially) remains uncertain even for experts, as screenwriter Goldman states in his well-known quote: "Nobody knows anything: Not a single person in the entire motion picture field knows for a certainty what's going to work. Every time out it's a guess and, if you're lucky, an educated one." (Goldman, 1983, p. 39).

Economists want to bring light into the dark and explain what makes movies successful and what characterizes successful movies. They understand movies as a bundle of characteristics and ask about the influence of individual characteristics. What influence do stars, advertising, genre, release date, movie length, reviews, and of course the budget have on the success of a movie? (for an overview see Chisholm et al., 2015). These studies acknowledge the extreme uncertainty mentioned in the quote and that "there are no formulas for success in Hollywood" (de Vany & Walls, 1999, p. 286). It is stated that experience-based studies are not instructive for rare events (de Vany & Walls, 1999, pp. 313-314) and that a separate economic model must be designed (Lieberman, 2006, p. 75), since the theoretical premises of economics do not hold (Baumol, 1986; Caves, 2002). Thus, as economists themselves note, what is needed is, first, a different theoretical approach in which movies are not understood as bundles of goods and, second, a different methodological approach because statistical evaluation of actual individual movie characteristics is not satisfactory.

Economic sociologists are less concerned with the success than with the market structure and consumers' related movie choice (Creton, 2009). To do so, they

take as their starting point the problems of action faced by market actors, which arise due to uncertainties about the product and the actions of fellow actors (Zuckerman & Kim, 2003, p. 33). The uncertainty for movie viewers is that they do not know whether a movie that is unknown to them will meet their subjective taste. Before the reasons for this are explained, it should be noted that the market players in production and distribution will also have to deal with this, because they too do not know in advance which movie will pay off financially. In institutional economics, institutions such as guarantees are discussed as solutions to situations of uncertainty (Akerlof, 1970). Such institutional arrangements can also be found between the studios and the distributors, since the practice of revenue sharing (Caves, 2003, pp. 79-80) minimizes the risk and allows the market to come into being. However, consumers do not receive any guarantees or the possibility of subsequent exchange and they are not offered to pay depending on the pleasure gained, which is why they are always confronted with a high degree of uncertainty. Such solutions seem to be unsuitable for the commodity movie, and this is probably also because once a movie has been made, it can be reproduced without significant costs in today's digital world. On the movie market, it is not the movie that is the scarce resource, but the consumer. So, the question for movie watchers is how to choose the right movie and not avoid this market because of frequent disappointments (which are possible due to the abundance of movies on the market). It is therefore important to find out what actors use as orientation when choosing a movie. Which influencing factors are relevant? This question is particularly relevant for going to the cinema, where the costs are higher than for in-home-viewing (entrance fees, invested time). To prevent a collapse of a market, Fligstein (2001, p. 17) argued, stable worlds are needed. Focusing on what problems market actors face and how this affects market structure is a promising approach for economic sociology. The interest of actors (suppliers as well as demanders) in the right choice of action and the uncertainty about it also raise the empirical question of which solutions are considered adequate.

The Economics of Singularities

Lucien Karpik (2010) offers an explanatory approach to how actors gain sufficient predictability of expectations to decide on a singular good. His framework explicitly refers to markets that cannot be understood by the standard model of economics, since the traded goods are characterized by three properties: multidimensionality, uncertainty and incommensurability. Multidimensionality means that the product has a structure that is composed of different properties. In contrast to the understanding that these bundles of characteristics are an accumulation of dimensions, as is common practice in economics (Lancaster, 1966), the individual qualities cannot, however, be evaluated individually (Karpik, 2010, pp. 24-26). Karpik himself uses the movie example to make it clear that the specific composition of script, actor,

music, camera angles, etc. forms a unity. The interplay of these individual dimensions, each of which brings its own qualities, results in the complex commodity of movie. Apart from this specific configuration of the qualities of a movie as such, it cannot be viewed in isolation from external circumstances. The audience, the quality of the copy and, for example, the seating comfort also determine the perception and evaluation of the movie and thus form further dimensions (Karpik, 2010, p. 39). Secondly, singular products are accompanied by uncertainty in two forms for consumers. Strategic uncertainty exists due to the different interpretations of a good by the seller and the consumer. Supplier and customer (generally different actors) may have different understandings of a product, which may lead to disappointment in the latter. Since the different possibilities of perception and interpretation are part of the nature of singular goods, there can be no guarantee that a product will be perceived in the same way. Different people may not have the same understanding of what constitutes a *good* movie. Treating quality uncertainty as a second form, Karpik describes the problem that at the time of purchase the buyer is not able to know whether he or she will be satisfied with the quality. The quality assessment is subject to a situation- and person-specific—and thus individual—assessment (Karpik, 2010, pp. 11-12, pp. 26-30). Only after watching a movie, it can be evaluated, prior to its viewing the quality remains uncertain. Uncertainty thus results from the fact that each individual actor has subjective and individual tastes that make one thing pleasing and another displeasing. This judgment of taste cannot be logically derived. The assumption about what will meet one's taste drives the choice of action (Arendt, 2003).³ Thirdly, incomparability determines singular goods. On the one hand—since, with reference to multidimensionality, one cannot speak of exactly the same thing due to the diversity of points of view—because every consumption is unique in its constellation, and on the other hand, because a plurality of value systems exist. The former means that each consumer perceives the same singular good from his or her individual perspective, which is influenced by subjective taste and situation, and the latter refers to the fact that there are no objective criteria on the basis of which a general ranking can be established. Although each person can compare something on his or her own to express preferences, these are not universally valid (Karpik, 2010, pp. 12-13). This means that even with the knowledge of certain characteristics of a movie (actors, length of the movie, etc.) each movie (and every viewing with its unique setting) is unique and must be evaluated individually. According to Karpik, these three product characteristics cause the previous economic models to fail. Likewise, de Vany and Walls come to the conclusion: “It is hard to imagine making choices in more difficult circumstances” (de Vany & Walls, 1999, p. 315).

3 Currently, the relevance of judgments and even their disparity among experts is prominently discussed by Kahneman et al (2021).

Consequently, according to Karpik, uncertainty in the demand for singular goods should not be understood as a risk that can be calculated. He argues that no knowledge allows an objective assessment of the probability of being satisfied and therefore the best choice, in a strict rational manner, cannot be made in advance (Karpik, 2010, pp. 35-43). Nevertheless, actors strive to make the right choice and therefore look for guidance that provides them with good reasons (Karpik, 2010, p. 67). They find this orientation in so-called *judgment devices* (Karpik, 2010, p. 44) that evaluate singular goods. Judgment devices can be institutions, persons, discourses, advertising messages, texts, and videos. They all convey knowledge and thus allow the consumer to make an educated choice, just like other *market devices* (Callon et al., 2007). In general, judgment devices make products visible in a market. Thus, Karpik (2010, p. 152-153) argues with regard to movies that smaller French movie productions are seen less often in cinemas precisely because judgment devices advertise them less. He assumes that the presentation and promotion of a movie by judgment devices is what attracts people to the cinema in the first place. The market for singular goods is therefore *embedded* in judgment devices. Karpik shows that in different markets, different judgment devices are socially valid and he distinguishes between different regimes, each with its own logic. The analysis of the relevant judgement devices is therefore interesting for economics, as they determine financial success, but also for sociology, which is interested in how culturally shaped beliefs affect the market structure. In line with the sociology of valuation and evaluation, different techniques and actors are distinguished with the aim of identifying different social mechanisms and their legitimacy in determining value and making judgements (Lamont, 2012). Which judgment devices are used in the movie market and what is the logic of market coordination? In order to answer this question, the various judgment devices on the movie market will be presented using Karpik's typology and subsequently hypotheses will be formulated based on these.

Karpik distinguishes five groups of judgment devices: the personal network and impersonal appellations, cicerones, rankings, and confluences. These are distinguished according to the nature of the knowledge provided, i.e., whether they qualify knowledge absolutely (substantial) or relative to others (formal). On the other hand, they are distinguished according to whether they intend to increase sales (commercial) or to enlighten the customer (critical). Karpik claims that commercial devices are increasingly found in large markets.

Networks

According to Karpik, personal contacts are particularly relevant when a personalized product is in demand. In this case, the trade network and the practitioner network are used. But even if a singular good is judged more by aesthetic criteria—as can be assumed with movies—actors can receive evaluations and information

about movies from family, friends, and colleagues, i.e., the personal network. The mostly orally acquired knowledge from such familiar persons offers some advantages, because on the one hand it can be obtained with little time and without further costs and on the other hand it can be classified as credible. The statements of the personal network are trusted (Karpik, 2010, pp. 183-185).

Sociology has long analyzed personal networks, and especially in the new economic sociology they are prominently researched in terms of information transfer and confidence building (Burt, 1992; Granovetter, 1973). One of the first studies to deal with the influence of personal contacts on going to the movies was the survey by Katz and Lazarsfeld (1964, p. 180), which showed the effectiveness of a personal recommendation and simultaneously its nonetheless rare consultation. Compared to recommendations from newspapers and magazines, personal recommendations have been much more effective in driving actual cinema attendance. The importance of personal ratings on cinema-going is pointed out by Faber and O'Guinn (1984). More recent studies on word of mouth also show that it has a high influence on movie success (Y. Liu, 2006).⁴

Appellations

The group of appellations includes product and umbrella brands, designations of origin, quality marks, or indications of professional qualifications. They are intended to signal certain characteristics and quality guarantees for a product (Karpik, 2010, pp. 45-46). In the case of movies, for example, this would be an indication of whether it is a movie from Hollywood or Bollywood, and the names of certain studios involved would also allow an assessment of the movie. Movie series such as 'James Bond' or 'Star Wars' are also included, as these labels allow many people to make an assessment and thus determine expectations. Studies show that movies attract a comparatively larger audience on the opening weekend if they are sequels (Moon et al., 2010, p. 114). Overall, sequels have greater commercial success (Ravid, 1999). The ratings of the movie templates or predecessors influence the demand for a future movie (Situmeang et al., 2014). These (brand) names are to be regarded as commercial entities, as they are used to attract customers and thus increase sales. Other appellations on the movie market include information on the genre (e.g., comedy, drama, or action) (commercial) and recommended age ratings (critical). The indication that stars are involved in the movie (actors, directors, etc.) can also be considered a commercial judgment device, since their name is perceived as a trademark (Levin et al., 1997, p. 177). Stars can be a guide for

4 Karpik's distinction between personal and impersonal devices is blurred with regard to the internet as a communication platform, which, as he himself notes, he has not taken into account (Karpik 2010, p. 131). For the distinction between traditional word of mouth, microblogging word of mouth, and electronic word of mouth see Hennig-Thurau et al. (2015).

viewers to assess in advance whether they will like the movie, but do not guarantee success. The numerous studies working with aggregated sales data, which investigate the question of whether the participation of stars influences the success of a movie, come to different conclusions (see in comparison A. Liu et al., 2014, p. 386; Boatwright et al., 2007, p. 410; Basuroy et al., 2003, p. 106).⁵ The studies are also criticized for their recourse to real data: “Movies are complex products [...] it is impossible to attribute the success of a movie to individual causal factors.” (de Vany & Walls, 1999, p. 285). Furthermore, the literature indicates that a star cast also has a positive influence on other judgment devices such as professional movie critics (Hennig-Thurau et al., 2012, p. 272). The judgment devices of the appellations group mentioned here have in common that they qualify substantially, since they work without the hierarchical comparison to competing movies.

Cicerones

The group of cicerones includes impersonal judgment devices that impart knowledge and present assessments. These include both general persons and critics but also products such as guides (Karpik, 2010, p. 46). For the movie market, two judgment devices can be identified that belong to this group. On the one hand, these are professional critics and, on the other hand, other consumers who also publish their movie reviews as critics, especially on the Internet. What both have in common is that they generally have no intention of increasing the demand for movies and that their reviews are not to be equated with rankings; they are critical and substantial.

Amateur critics are discussed in the literature under the terms ‘electronic word of mouth’ (Hennig-Thurau et al., 2015), ‘online consumer review’ (Mellet et al., 2014), and ‘user-generated content’ (Gopinath et al., 2013). On websites (e.g., ‘IMDb.com’), users share reviews among themselves and assign a summarizing score. Movies that have received several reviews are usually given an average value of the points awarded. As with the personal network, access to such virtual networks is easy and questions can be answered readily (Bialecki et al., 2017). However, trust in these online communities is not based on personal relationships, but results from the mass of ratings, which Mellet et al. (2014) calls democratization of markets. Studies have shown that the trust placed in these evaluations is high if they are not only based on many judgements but also differ only slightly from each other (Ji et al., 2015).

In contrast, a critic is defined as “a person usually employed by newspapers, television stations or other media who screen newly released movies and provide their subjective views and comments on the movie for the public’s information.” (Cones, 2013, p. 99) They possess expertise and are attributed neutrality (Hennig-

5 These studies fail to provide a realistic explanation because of the tautological definition: if a person is a star when he or she attracts audiences and a large audience determines film success, then by definition a star indicates film success.

Thurau et al., 2015, p. 377). They have a special position because they are usually the first to be allowed to watch movies, conduct interviews, and report on them (Eliashberg & Shugan, 1997; Ravid et al., 2006). In addition, not only do they generate a reputation, but they themselves depend on a high one to be consulted, which is why it is assumed that critics will judge as they expect their colleagues to do (Ravid, 1999, p. 489). A large number of studies examine the question of the importance of critics on the movie market (for an overview see Hennig-Thurau et al., 2012). It is unclear whether they influence or foresee movie success (Eliashberg & Shugan, 1997). Gemser et al. (2007) argue on the basis of their analysis of the Dutch movie market that the consultation of critics depends on the type of movie: critics influence art house viewers but not blockbuster viewers. For this purpose, newspaper reviews and the box-office receipts of the respective movies were correlated, which, however, does not allow any conclusions to be drawn about causality, since it remains unclear whether the cinema-goers had even consulted these reviews and whether they were guided by other judgements. Without differentiating between these two sub-markets in a factorial survey design, Tsao (2014) comes to the conclusion that consumer criticism has a greater influence than expert criticism. This result was corroborated in an evaluation of databases (Kumar et al., 2016).

Rankings

Rankings compare one or more criteria to create a hierarchical order, which by definition makes them formal devices. Karpik (2010, p. 46) distinguishes two categories: expert rankings and buyers rankings.

Expert rankings result from the importance of awards, which are given by a jury consisting of authorities. In the case of movies, as in the case of music and literature, these are awarded annually at festivals. By winning prizes (e.g., a Golden Globe or Oscar) in different categories (best movie, best actor, best supporting actor), singular goods receive a higher status. Originally, festivals set themselves apart from the economy. They concentrated on art house movies and, with their focus on artistic aspects, formed an antithesis to commerce, especially from Hollywood (de Valck, 2007). Some studies have shown that rankings correlate positively with movie success (R. A. Nelson et al., 2001; Zhuang et al., 2014), although the methods do not allow any conclusions to be drawn about causality. Since these rankings are not created to boost the sales of certain products and are based on aesthetic criteria (Simonton, 2004), they are to be regarded as critical judgment devices.

Buyers rankings are based on the criterion of sales numbers. The more movie-goers a movie has, the higher its position in the charts. This results in a clear order, whereby the list presented in the media usually only mentions the top 5 or top 10. The influence of such charts on the movie market has not yet been empirically investigated. When cinema-goers visit a movie simply because they are driven by

its success, a self-reinforcing process occurs, as Merton (1968) described with the Matthew effect—a social phenomenon observed in many contexts. In fact, however, this is not the intention of such rankings. Since they provide factual information about an aspect, they are also regarded as critical judgment devices.

Confluences

Finally, under confluences, Karpik includes a wide variety of sales techniques designed to encourage purchases, from window displays to the various types of advertising (Karpik, 2010, p. 46). By definition, confluences are therefore commercial judgment devices that usually work with substantial knowledge.

Advertising plays a major role in the movie market: in newspapers, on the radio, on television, with posters, and in the cinema itself, attempts are made to make movies appealing (H. Liu, 2016). A special instrument in this context are trailers in which excerpts from the movie are used to give a foretaste (Creton, 2009, pp. 146-147). Several studies show that advertising promotes sales, although the correlation does not increase proportionally with the budget. In the literature, advertising is therefore understood as a multiplier that is effective in combination with cicerones and rankings (Basuroy et al., 2006, p. 287; Gopinath et al., 2013; Hennig-Thurau et al., 2012, p. 270).

Coordination Regimes and Derived Hypotheses

As Table 1 summarizes, there are various judgment devices in the market for motion pictures. Which of these many different judgment devices are used by cinematographic viewers? According to Karpik (2010, pp. 96-105), different judgment devices are used for the various singular goods, which differ in their objective and nature of the knowledge. Thus, different logics prevail in the various markets, of which Karpik identifies seven. In the following, Karpik's reflections on the coordination regimes will be pursued in order to derive hypotheses on which judgment devices are used in the selection of motion pictures.

Karpik's classification of the coordination regimes is based on a number of plausible examples. The distinctions are made inductively, which is why the approach has been criticized (Hutter, 2011a, p. 792). Karpik (2010, pp. 152-157) himself uses the example of the cinema market as an illustration and also raises the question of how a movie is chosen here. However, he does not answer this question empirically, but relies on moviegoers' self-reported frequency of cinema visits in order to form assumptions about which judgment devices are relevant for them. His main aim is to show that judgment devices are necessary for market coordination. He states that "[t]he movie market is hybrid: it comes under both the authenticity regime and the mega regime." (Karpik, 2010, p. 156) In the following, this thesis will be considered and examined in a more differentiated manner. Based on the

Table 1 Judgment devices in the movie market

Groups and judgment devices	Examples	Objective	Knowledge
Networks			
personal network	family, friends and acquaintances	–	–
Appellations			
designation of origin	Hollywood, Bollywood	commercial	substantial
brand name	Walt Disney, Paramount	commercial	substantial
star participation	Steven Spielberg, George Clooney	commercial	substantial
Cicerones			
expert critics	in the media	critical	substantial
laymen critics	on the internet	critical	substantial
Rankings			
expert rankings	Oscars, Golden Globes	critical	formal
buyers rankings	charts	critical	formal
Confluences			
spatial organization	presentation of the cinema	commercial	substantial
advertisement	in the media, film trailer	commercial	substantial

results on market coordination by Gemser et al. (2007) and Holbrook and Addis (2008), it is assumed that the authenticity regime is found in art house cinema and the mega regime in blockbuster cinema. Accordingly, the preference for certain movies should be accompanied by the use of specific judgment devices.

Both coordination regimes have in common that the market is mainly characterized by impersonal and substantial devices. However, while critical devices play a more important role in the authenticity regime, more commercial devices are found in the mega regime (Karpik, 2010, p. 165). Karpik also makes assumptions about consumer commitment. He assumes that the consumers of the authenticity regime are characterized by autonomy, “the capacity to define and maintain one’s personal tastes,” whereas the consumers of the mega regime are characterized by heteronomy and “accept the tastes embodied by the devices and/or products” (Karpik, 2010, p. 104). Based on this, the following hypotheses are formed.

Since consumers with a preference for art house movies are more strongly oriented towards their own tastes and critical instances, they should consult expert critics (hypothesis 1a) and expert rankings (hypothesis 1b) more than blockbuster viewers. In contrast, consumers with a preference for blockbuster movies should orient themselves more towards mainstream taste and be influenced by commercial devices. It is assumed that in the group of cicerones the laymen critics (hypothesis

2a) and in the group of rankings the charts (hypothesis 2b) have a greater influence on blockbuster moviegoers than on art house watchers. It is also assumed that the star participation has a higher influence on them (hypothesis 2c).

In addition, two further general hypotheses based on assumptions from the economics of singularities will be tested. First, according to Karpik (2010, p. 131), the personal network is never completely switched off. Regardless of which movies are preferred, it is assumed that the personal network is the most influential (hypothesis 3). This is justified by the fact that familiar people provide helpful knowledge (individually adapted) and are classified as trustworthy (no opportunism is assumed) (Karpik, 2010, p. 183, 2010, p. 65). On the other hand, according to Karpik (2010, p. 124), market competition for singular goods is more about quality than about price. Since customers attach more importance to the quality of a product than to its price, price should not be the most relevant orientation criterion (hypothesis 4).

The Empirical Investigation of the Choice of Movie

Methods and Data

The factorial survey is, it will be argued here, a suitable method because it has two necessary properties. First, the experimental design allows to identify causality and the influence of individual factors on an evaluation. As will be shown with the help of a comparison, correlations can be revealed that cannot be determined by a classic survey alone. Second, the factorial survey allows us to control for the social and complex world in order to identify the influence of specific factors. As has been shown, the complexity of the real world makes it difficult to identify influencing factors, as the tastes of the actors distort the results. Since it is not taste preferences that are to be investigated, but rather the strength of influence of the market structure determining judgment devices, it can be seen as an advantage to work with abstract vignettes that exclude the disturbance variable of taste. For example, with the factorial survey it is possible to determine the influence of a star's participation without attaching it to specific stars. Therefore, those judgment devices are included in the factorial survey for which the influence can be determined without in fact collecting taste preferences. Judgment devices for which this is not possible are therefore not included in the factorial survey. This includes the group of confluences because advertising predominantly draws attention to movies and tastes are either directly affected or not, which is why a separation is not feasible. The same applies to the judgment devices brand name and designation of origin. Here, too, a meaningful query as to whether an exemplar has an influence is not possible without naming specific brands or designations of origin at the same time. In other

words: With these (commercial) judgment devices, positive and negative levels cannot be identified without recourse to taste preferences.

For the data collection, a factorial survey was implemented in an online questionnaire. The goal of the experimental design is to determine the causal influence of judgment devices in order to test the formulated hypotheses. In addition to the vignettes, a classical questionnaire with item queries preceded the vignettes. Social-demographic information was requested and it was also asked which of the mentioned judgment devices influence the personal choice of movie, which allows a comparison of the item survey and the factorial survey as well as a more in-depth analysis. To test the hypotheses about the different movie type preferences, it was asked whether blockbuster or art house cinema was preferred or whether a clear assignment was not possible (indifferent). For the factorial survey the participants were asked to put themselves in the scenario that the following ratings and characteristics about a cinema movie are available to them and they were asked: "How likely would you go and see this movie?" (see Figure 1) Table 2 summarizes the dimensions and factor characteristics of the factorial survey. Each vignette contains seven dimensions, which can take one of two levels: a positive and a negative one. The values are chosen in such an abstract way that they do not reveal any taste preferences. For expert and laymen critics, a realistic star rating was used to avoid monotonous vignettes. No extreme values were chosen, as this is more in line with actual ratings. The admission price was based on the upper and lower quintile of the admission price to German cinemas.⁶

Since all 128 possible combinations were included in the study, it is a complete 2^7 design. The vignette universe is divided into 16 sets, so that each study participant is presented with 8 different vignettes. The composition of the sets is targeted to avoid confounding effects and to ensure that each set contains a maximum of different combinations of factor levels. This should also ensure that all vignettes are processed with approximately the same frequency. To achieve this, the online questionnaire was programmed in such a way that the random assignment of a respondent to a set takes into account an equal distribution across all sets. The 11-step answer scale for each vignette ranges from "0 – extremely unlikely" to "10 – extremely likely." These vignettes are shown to the participants one after another and on separate pages after a classic online survey. They were allowed to jump back to previous statements and adjust the answers. The individual vignettes were presented in tabular form, so that the dimensions are always in the same order on the left-hand side and one of the two characteristics is always shown on the right-hand side. Thus, participants were enabled to make a quick and uncomplicated com-

6 In Germany, the entrance fee for cinema movies depends on the film and the day of the week. In countries such as the USA and Australia, each film is offered at the same time for the same admission price, which is discussed in literature under the term 'movie puzzle' (Chung 2015).

For the following eight movies, please indicate how likely it is that you would see these movies in the cinema.

Characteristics that concern subjective taste are deliberately neglected in these abstract film characteristics. Put yourself in the scenario that these movies have equally aroused your interest and that you are now aware of the information given.

How likely would you go and see this film?

The film ...

is ... by family or friends.	recommended
has a star in it.	yes
is rated by experts	2 out of 5 stars
is rated by laymen	4 out of 5 stars
has won awards.	no
is in the top five of the charts.	yes
costs an entrance fee of	12 €

extremely unlikely [0 1 2 3 4 5 6 7 8 9 10] *extremely likely*

Figure 1 Instructions for answering the vignettes and example vignette

Table 2 Dimensions of the factorial survey

Judgment device	Dimension: The film ...	Positive level	Negative level
personal network	is ... by family or friends.	recommended	not recommended
star participation	has a star in it.	yes	no
expert critics	is rated by experts	4 out of 5 stars	2 out of 5 stars
laymen critics	is rated by laymen	4 out of 5 stars	2 out of 5 stars
expert rankings	has won awards.	yes	no
buyers rankings	is in the top five of the charts.	yes	no
entrance fee	costs an entrance fee of	6 €	12 €

parative assessment. The presentation as a table and the maintenance of a uniform ranking of the dimensions has the goal of making it easy for the study participants to answer in order to prevent study dropouts or even rash decisions. While the table presentation is not considered more problematic than a text format presentation in the literature, there is, however, some evidence that the constant ranking of the

dimensions could lead to response heuristics. Even though such effects are only suspected for complex vignettes with twelve dimensions or more, they cannot be completely ruled out for this study (Auspurg & Hinz, 2015, pp. 70-72; Auspurg & Jäckle, 2017; Sauer et al., 2020).

The participation in the online questionnaire was enabled for students of the local universities via an email distribution list. A raffle of vouchers for online shops was used to motivate them to participate. The data collection took place over two weeks in January 2017.

Results

In total, the questionnaire was started 994 times and completed 910 times, although not all questions were always answered. The average processing time is 7.5 minutes. The gender ratio is balanced with approx. 50.7% female and 47.7% male participants. The average age of the participants is 26 years. 62.9% prefer blockbusters, 10.7% art house and 26.4% say they are indifferent.

A total of 7278 evaluations are distributed among the 128 vignettes, whereby the answer scale was fully exhausted. Each vignette was rated between 54 and 60 times. However, this almost equal distribution is not evident with regard to preference groups: 2 of the 16 sets were answered by only 2 persons who also stated that they preferred art house cinema. However, this is not seen as problematic due to the orthogonal design.

The statistical test procedure used is the single factor analysis of variance (ANOVA), which can be considered as a special form of regression analysis (Rutherford, 2001, p. 9).⁷ The evaluation includes all judgements of the persons who, in addition to the vignette evaluations, also made a self-classification into one of the three groups. The sample thus comprises 905 persons, of which 568 persons (62.8%) belong to the group Blockbusters, 100 persons (11%) to the group art house and 237 persons (26.2%) to the indifferent group. Table 3 shows the results of the ANOVA with the regression coefficient B, the significance level, and the partial η^2 as an effect strength measure in four complete models; the first model includes all consumer groups and the other three reflect the results of the individual consumer groups.⁸ The inclusion of gender and age as moderator variables does not produce any significant effects, so that these third variable effects (e.g. because art house movies are preferred by female participants) can be excluded. With a corrected coefficient of determination above 0.3, all models are of high quality for the social

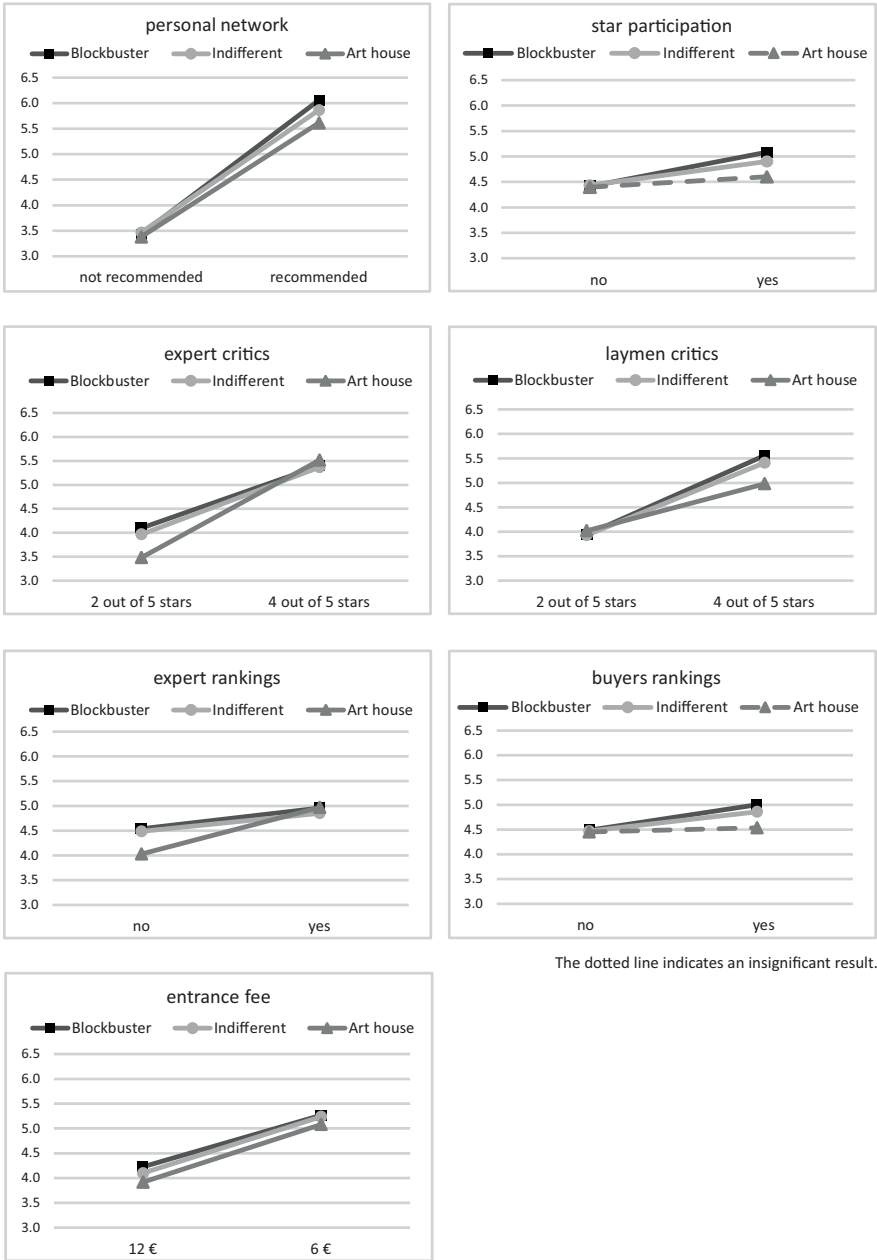
7 A hierarchical-linear model yields almost exactly the same results (Schoppek, 2015; Steiner and Atzmüller, 2006, p. 138).

8 Because of the orthogonal data matrix, the standard errors are the same for all variables in a model (Blockbuster, art house and indifferent: 0.052; Blockbuster: 0.065; Art house: 0.157; Indifferent: 0.108).

Table 3 Single factorial analysis of variance

	Blockbuster, art house and indifferent ($\emptyset = 4,714$) n = 905		Blockbuster ($\emptyset = 4,747$) n = 568		Art house ($\emptyset = 4,502$) n = 100		Indifferent ($\emptyset = 4,668$) n = 237	
	Regression coefficient B	partial η^2	Regression coefficient B	partial η^2	Regression coefficient B	partial η^2	Regression coefficient B	partial η^2
(constant)	0.684 ***	.008	0.662 ***	.011	0.691 **	.012	0.852 ***	.016
personal network	2.518 ***	.241	2.619 ***	.266	2.233 ***	.204	2.397 ***	.207
star participation	0.565 ***	.016	0.670 ***	.023	0.206	.002	0.466 ***	.010
expert critics	1.405 ***	.090	1.297 ***	.081	2.030 ***	.175	1.400 ***	.082
laymen critics	1.510 ***	.102	1.617 ***	.121	0.960 ***	.045	1.481 ***	.091
expert rankings	0.462 ***	.011	0.419 ***	.009	0.939 ***	.043	0.364 ***	.006
buyers rankings	0.431 ***	.009	0.510 ***	.014	0.083	.000	0.385 ***	.007
entrance fee	1.078 ***	.055	1.039 ***	.054	1.164 ***	.065	1.138 ***	.056
corrected R ²	.377		.402		.378		.339	

*** p < .001, ** p < .01



The dotted line indicates an insignificant result.

Figure 2 Profile diagrams: mean values of the likelihood of going to the cinema of the positive and negative levels, differentiated by movie preferences and dimensions

sciences.⁹ The interpretation of the effect size, indicated by the partial η^2 , is based on Cohen (1992): values smaller than 0.06 show small effects, values between 0.06 and 0.14 are interpreted as medium effects and larger values as strong effects. The effect size is given to easily assess the influence of a variable independent of its scale. The partial η^2 offers the advantage of allowing comparison to the same variable in other studies where other levels of measurement, covariates, or other factors are included. Overall, the mean value of the vignette assessments is 4.71; the standard deviation is 2.83. The analysis of variance indicates a significant difference in the distribution of variance between blockbuster and art house viewers: $F(1, 5340) = 4.426, p = 0.035$.¹⁰ These differences will now be examined along the hypotheses. For illustration and further interpretation, the estimated values of the positive and negative values for the individual dimensions are compared in profile diagrams in Figure 2.¹¹

The preference for art house or blockbuster cinema determines the influence of different judgment devices. The influence of expert critics on the art house group is rated as high with an effect strength of 0.175. The probability to visit a movie increases by about 2 units on the 11-level scale, if the movie receives 4 out of 5 stars instead of 2 from experts. As the influence of the variable on the group blockbusters is lower, hypothesis 1a is supported. The influence of the variable on the blockbuster group is also significant and can be classified as medium. The profile diagram illustrates that if experts' ratings were positive, all consumer groups would be more or less equally likely to visit the movie, whereas a negative rating would have a greater impact on the art house group. Hypothesis 1b is also supported, since the same phenomenon can also be observed with expert rankings. While awards have the least effect on the blockbuster group, and this is marginal ($\eta^2 = 0.009$ and about 0.4 scale units), it is higher in the art house group ($\eta^2 = 0.043$ and about 0.9 scale units). As hypothesized, laymen critics (hypothesis 2a) and buyers rankings (hypothesis 2b) have a higher impact on the blockbuster group. In the case of a negative rating from other consumers, the cinema visit probability for all groups is on average 4, but in the case of a positive rating it is about 1.6 units higher for the blockbuster group and almost 1 unit higher for the art house group. The influence is present and significant for all groups. Whether a movie is in the charts has no significant effect on the art house group ($\eta^2 = 0.0$), but has a slightly higher effect on the Blockbuster group ($\eta^2 = 0.014$) than expert rankings. A comparable picture can

9 However, Snijders and Bosker (1999, p. 99) point out that measures of determination in experimental procedures should be interpreted with caution and should not be given high priority.

10 Both the variance analyses of the groups blockbuster and indifferent, $F(1, 6438) = 0.803, p = 0.370$, and those between the groups art house and indifferent, $F(1, 2692) = 1.702, p = 0.192$, are not significant.

11 The boxplots can be found in figure 4 in the Appendix.

be seen in the variable star participation. The star participation has a positive effect on the blockbuster group ($\eta^2 = 0.023$), whereas it is not significant for the art house group ($\eta^2 = 0.002$). Hypothesis 2c is therefore also supported, although the influence of the variable is small. Hypothesis 3, according to which the network is the most influential judgment device, is also corroborated. Across all models, the recommendation or dissuasion by family and friends to visit a movie has the strongest effect ($\eta^2 = 0.241$). There are two aspects to consider that can condition the strong influence. First, going to the movies is a social event for most people. A movie is attended with others, which is why the opinions of these others can be very influential. Second, one reason for the strength of the network effect may be that this factor was listed first in all vignettes. As eye-tracking studies have shown, more attention is paid to the first statements (Galesic et al., 2008). Potentially, this order effect plays a role despite the low complexity of the vignettes. Finally, hypothesis 4 can also be supported. Nevertheless, the prices given here have a significant and noteworthy influence on the probability of going to the cinema ($\eta^2 = 0.055$). This may also be due to the fact that students were surveyed, who have to keep a closer eye on their expenses due to their presumably tight budgets. Overall, the results are not generalizable. As in all experimental designs, external validity is secondary to internal validity (Auspurg & Hinz, 2015, p. 62). The general model as well as the group-specific ones make it clear that the two formal devices, charts and awards, have the least relevance, which speaks for a coordination based on originality and specific knowledge.

The variables used in the hypotheses, with the exception of price, were also used with some other previously mentioned judgment devices in a classical item query. Figure 3 shows for the three consumer groups how often it was stated that the respective judgment devices are relevant for the decision. The results of the item query support hypotheses 1 to 3.

However, the item query does not reveal any effect strengths, which is why the factorial survey offers considerable added value. Thus, it is surprising that some variables of the blockbuster group and the art house group are almost equal, which can only be viewed in a more differentiated manner thanks to the vignette analysis. A second surprising finding is that the two variables 'star participation' and 'buyers rankings', which were not significant in the vignette analysis, are nevertheless accepted in the item survey. A possible interpretation for this result is that these variables are indeed classified as relevant for decision making but not in the assumed sense for art house watchers. For some of the art house group, the fact that a star is involved or that the movie is in the charts may be a deterrent. Perhaps some art house fans reject on principle the star presence typical of blockbusters. Or maybe it's certain stars that make people go to the movies. This cannot be determined from the data collected but could be the subject of future studies. Nevertheless, it must be noted that factorial surveys and classic item queries can complement

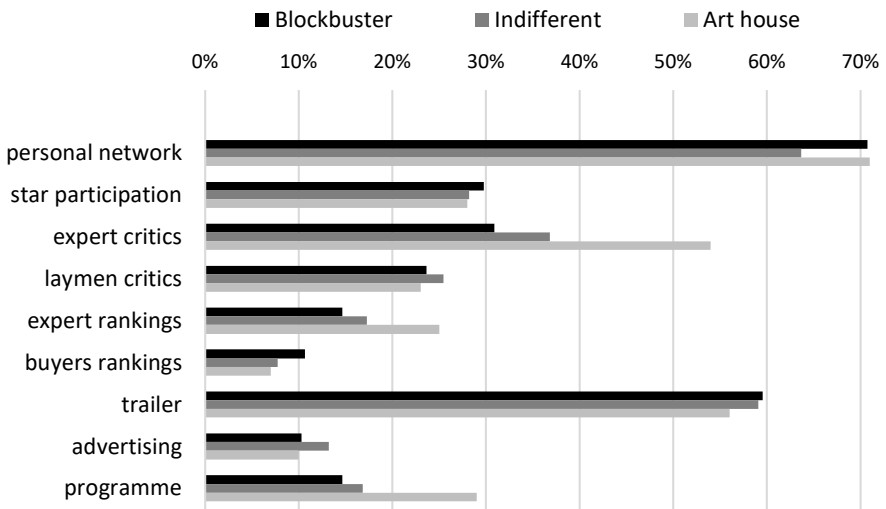


Figure 3 Item query: judgment devices relevant for choice

each other. Finally, the item survey shows that trailers are assigned a high decision-making significance by all consumer groups. Trailers are regarded as helpful judgment devices, which are nevertheless not as decisive as the personal network, thus asserting hypothesis 3.

Conclusion and Discussion

The choice of a movie confronts consumers with a high degree of uncertainty, as they do not know in advance whether they will like the selected movie. This market was used to investigate what moviegoers look for when deciding to watch a movie. To find this out, it was argued, the factorial survey offers some advantages. Karpik (2010) offers an explanatory approach for the market coordination of such special goods. The approach was positively received in the academic world, but its considerations were rarely empirically tested, which Karpik himself justifies with the difficulty of the research subject. He states that “it is not only a matter of increasing the number of markets; more specific data and, as a result, more elaborated analysis are also needed.” (Karpik, 2010, p. 256). The pursuit of the research question presented in this paper is therefore also intended to examine the extent to which factorial surveys are useful in dealing with problems of uncertainty management in complex situations. This also contributes to the debate on experimental methods in economic sociology (Beckert & Streeck, 2008; Keuschnigg & Wolbring, 2015;

Kittel, 2015; Wolbring, 2017) by listing the advantages and disadvantages of factorial surveys in research on coping with uncertainty in particular (cultural) markets (Watts & Salganik 2011).

Karpik (2010) argues that market players are not paralyzed because they find orientation in so-called judgment devices to identify the right product. In different markets different judgment devices are used, resulting in different coordination regimes. In line with this argumentation, it was assumed that the market for motion pictures is divided into two coordination regimes. Previous research has argued that blockbuster cinema should be distinguished from art house cinema. The former follows the logic of the mega regime and the latter the logic of the authenticity regime. The established hypotheses were tested with a factorial survey embedded in an online questionnaire and compared to the classical item-questionnaire. The results significantly support all hypotheses on the economics of singularities and the differences between the two groups of consumers. Thus, it can be concluded that Karpik's approach is suitable for the development and testing of empirically testable hypotheses. The result is that potential moviegoers use different judgment devices depending on their taste when deciding whether and which movie to watch. The distinction as to whether judgment devices are formal or substantial respectively critical or commercial helps to distinguish which devices are socially valid; however, his approach leaves unexplained *why* a particular judgment device is chosen. Cinema-goers who prefer art house movies as well as those who prefer blockbuster movies orientate themselves particularly towards the personal network. While the former, however, appreciate expert judgements (critics and awards), the latter tend to use the judgements of other consumers and also charts. The results presented complement the state of research on demand in the movie market and show causality.

Not all of the influencing factors mentioned in the second section could be used as variables in the factorial survey for two reasons. On the one hand, a vignette should not be too large and complex to avoid heuristic response behavior and to achieve meaningful results. A rule of thumb is to use approximately seven dimensions (plus or minus two) (Auspurg & Hinz, 2015, pp. 18-19). On the other hand, dimensions are not suitable for use in factorial surveys if they cannot be implemented in a meaningful way. With the exception of star participation, this applies especially to commercial devices. Firstly, this is due to the fact that advertising is usually not consciously consulted and has a subconscious effect, and secondly, because an implementable use of factors does not measure the influence of a judgment device but rather taste preferences. This is the case, for example, when different brand names are given. Contrary to critical devices, which provide positive and negative information, the aim of commercial entities is to promote sales, which is why a product is only presented in a positive light – whether the actor judges this presentation (e.g., trailer or advertisement) positively or negatively depends on his

or her subjective taste, which is deliberately not collected. And of course, other aspects, such as the particular content of the movie, play a significant role in the choice of movie in real life, which was neither intended nor could be addressed here. These alleged disadvantages of the method therefore have significant advantages for the objective pursued here. In contrast to work with real data, a more abstract situation can be constructed with the experimental design, in which taste preferences and other social effects such as status and reputation do not distort the influence of judgment devices. These disturbing factors cannot be eliminated in the real world, especially in the case of singular goods that are characterized by complexity. "Indeed, all experiments eliminate much of the 'noise' of real-world settings. However, this is the key to ensuring high internal validity; hence, it is a strength rather than a limitation." (Auspurg & Hinz, 2015, p. 115). And the advantage over classical questioning is not only that causality can be identified, but also the strength of its effects. It was shown that vignette surveys and item queries complement each other well, as results are compared and thus misinterpretations can be avoided. Experimental design is thus particularly fruitful in the analysis of action-guiding factors in complex, uncertain situations. The scenarios of the vignettes have to be chosen with care and critically examined regarding their factual content. Thus, the issue remains that in the study presented here, the specified variables are not always available or sought out, and opinions of judgment devices are of course also more complex and greatly simplified in the investigation. Methodologically, when working with factorial surveys, it is important to keep in mind that they always reduce the complexity of the real world. This is a great benefit for some research subjects and objectives. But even if this property should not be the main reason for choosing the method, it should be taken into consideration.

Taste has been identified as a confounding factor that makes it difficult to measure the influence of judgment devices in the real world. The problem results from the complexity of the singular good movie and the different taste preferences of its consumers. So how are taste preferences controlled? The experimental design allows for the control of confounding factors with two techniques that were used here. On the one hand, an abstract situation was constructed in which aspects of taste were *eliminated*. That means, only judgment devices were included that allow an evaluation without addressing issues of taste, and for these the dimensions were also chosen to exclude aspects of taste. As noted, this is a difficult task that may not have worked consistently because, for example, the mere fact that a star is involved in the movie influences taste perception. An upstream query could help to find out what issues of taste exist, so that these can be better controlled. On the other hand, the preferred type of movie was queried and so, according to this self-report, the taste preference could be kept *constant* during the analysis. Keeping confounding factors constant becomes useful when it is not possible to eliminate them. This is the case with the global preference for a certain type of movie. This global prefer-

ence should not have changed for the evaluation of several vignettes, and as has been shown, differences between the two groups can thus be discerned.

Finally, the question of whether the factorial survey design chosen here is appropriate for the question pursued here needs to be critically addressed. As noted in section 1, the question of the correct choice of movie can be considered in different settings. The factorial survey was used to find out what influence different judgment devices have on whether a movie is perceived as the proper choice and consequently whether the movie theater is visited. The dependent variable is therefore the probability of going to the cinema, which was not asked dichotomously as a yes/no statement but on a multilevel scale to provide a more precise picture. In this setting, the decision to go to the cinema must first be made. A different setting exists when the decision to go to the cinema has already been made and one of the movies on offer now needs to be selected. This may be the case, for example, when friends have arranged to go out to the cinema together. In this case, the social gathering would be the main goal and the choice of the right movie would be secondary and shaped by this context. In such a case, an alternative survey design would be appropriate. A *discrete choice experiment* would be suitable in which two or more movies are contrasted and the respondents are asked to choose one of them. In order to achieve a high degree of validity, it is important to ensure that the experimental design is conceived in such a way that it corresponds as closely as possible to the real-world situation. For the area of concern here, therefore, it would have been necessary to find out how the movie-going situation is set up in reality. It seems plausible that different types of moviegoers can be identified in this respect as well.

References

- Akerlof, G. A. (1970). The Market for "Lemons": Quality Uncertainty and the Market Mechanism. *The Quarterly Journal of Economics*, 84(3), 488–500.
- Arendt, H. (2003). *Responsibility and Judgement* (J. Kohn, Ed.). Schocken.
- Aspers, P. (2016). *Orderly fashion: A sociology of markets*. Princeton University Press.
- Auspurg, K., & Hinz, T. (2015). *Factorial survey experiments*. Sage.
- Auspurg K, Jäckle A. (2017). First Equals Most Important? Order Effects in Vignette-Based Measurement. *Sociological Methods & Research*, 46(3), 490-539.
<https://doi.org/10.1177/0049124115591016>
- Basuroy, S., Chatterjee, S., & Ravid, S. A. (2003). How Critical Are Critical Reviews? The Box Office Effects of Movie Critics, Star Power, and Budgets. *Journal of Marketing*, 67, 103–117.
- Basuroy, S., Desai, K. K., & Talukdar, D. (2006). An Empirical Investigation of Signaling in the Motion Picture Industry. *Journal of Marketing Research*, 43(2), 287–295.
<https://doi.org/10.1509/jmkr.43.2.287>
- Baumol, W. J. (1986). Unnatural Value: Or Art Investment as Floating Crap Game. *The American Economic Review*, 76(2), 10–14.

- Beckert, J. (1996). What is sociological about economic sociology? Uncertainty and the embeddedness of economic action. *Theory and Society*, 25, 803–840.
- Beckert, J., & Streeck, W. (2008). *Economic Sociology and Political Economy: A Programmatic Perspective* [MPIfG Working Paper 08/4]. Max-Planck-Institut für Gesellschaftsforschung. <http://www.mpi-fg-koeln.mpg.de/pu/workpap/wp08-4.pdf>
- Benjamin, B. A., & Podolny, J. M. (1999). Status, Quality, and Social Order in the California Wine Industry. *Administrative Science Quarterly*, 44(3), 563–589. <https://doi.org/10.2307/2666962>
- Bialecki, M., O’Leary, S., & Smith, D. (2017). Judgement devices and the evaluation of singularities: The use of performance ratings and narrative information to guide movie viewer choice. *Management Accounting Research*, 35, 56–65. <https://doi.org/10.1016/j.mar.2016.01.005>
- Boatwright, P., Basuroy, S., & Kamakura, W. (2007). Reviewing the reviewers: The impact of individual movie critics on box office performance. *Quantitative Marketing and Economics*, 5(4), 401–425. <https://doi.org/10.1007/s11129-007-9029-1>
- Burt, R. S. (1992). *Structural holes: The social structure of competition*. Harvard University Press.
- Callon, M., Millo, Y., & Muniesa, F. (Eds.). (2007). *Market Devices*. Sociological review monographs. Blackwell.
- Campbell, J. L. (2010). Valuing the Unique: The Economics of Singularities. *Administrative Science Quarterly*, 55(4), 683–685.
- Caves, R. E. (2002). *Creative Industries: Contracts between Art and Commerce*. Harvard University Press.
- Caves, R. E. (2003). Contracts between Art and Commerce. *Journal of Economic Perspectives*, 17(2), 73–83.
- Chisholm, D. C., Fernández-Blanco, V., Abraham Ravid, S., & David Walls, W. (2015). Economics of motion pictures: The state of the art. *Journal of Cultural Economics*, 39(1), 1–13. <https://doi.org/10.1007/s10824-014-9234-1>
- Chung, H. S. (2015). A Note on Uniform Pricing in the Motion-Picture Industry. *Hitotsubashi Journal of Economics*, 56, 231–242. <https://doi.org/10.15057/27597>
- Cohen, J. (1992). A Power Primer. *Psychological Bulletin*, 112(1), 155–159. <https://doi.org/10.1037/0033-2909.112.1.155>
- Cones, J. W. (2013). *Dictionary of movie finance and distribution: A guide for independent moviemakers*. Algora Publishing.
- Creton, L. (2009). *Économie du cinéma: Perspectives stratégiques* (4th ed.). Armand Colin.
- de Valck, M. (2007). *Movie festivals: From European geopolitics to global cinephilia*. Amsterdam University Press.
- de Vany, A., & Walls, W. D. (1999). Uncertainty in the Movie Industry: Does Star Power Reduce the Terror of the Box Office? *Journal of Cultural Economics* (23), 285–318.
- Dekker, E., & de Jong, M. (2016). What do book awards signal? An analysis of book awards in three countries. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2817852
- Eliashberg, J., & Shugan, S. M. (1997). Movie Critics: Influencers or Predictors? *Journal of Marketing*, 61, 68–78.
- Espeland, W. N. (2011). Lucien Karpik Valuing the Unique: The Economics of Singularities. Princeton, Princeton University Press, 2010. *Socio-Economic Review*, 9(4), 794–800. <https://doi.org/10.1093/ser/mwr010>

- Faber, R. J., & O'Guinn, T. C. (1984). Effect of Media Advertising and other Sources on Movie Selection. *Journalism & Mass Communication Quarterly*, 61(2), 371–377. <https://doi.org/10.1177/107769908406100219>
- Fligstein, N. (2001). *The architecture of markets: An economic sociology of twenty-first-century capitalist societies*. Princeton University Press.
- Gadrey, J. (2008). Regards croisés sur L'économie des singularités de Lucien Karpik. *Revue Française De Sociologie*, 49(2), 379. <https://doi.org/10.3917/rfs.492.0379>
- Galesic, M., Tourangeau, R., Couper, M. P., & Conrad, F. G. (2008). Eye-Tracking Data: New Insights on Response Order Effects and Other Cognitive Shortcuts in Survey Responding. *Public Opinion Quarterly*, 72(5), 892–913. <https://doi.org/10.1093/poq/nfn059>
- Gemser, G., van Oostrum, M., & Leenders, M. A. A. M. (2007). The impact of movie reviews on the box office performance of art house versus mainstream motion pictures. *Journal of Cultural Economics*, 31(1), 43–63. <https://doi.org/10.1007/s10824-006-9025-4>
- Goldman, W. (1983). *Adventures in the screen trade: A personal view of Hollywood and screenwriting*. Abacus.
- Gopinath, S., Chintagunta, P. K., & Venkataraman, S. (2013). Blogs, Advertising, and Local-Market Movie Box Office Performance. *Management Science*, 59(12), 2635–2654. <https://doi.org/10.1287/mnsc.2013.1732>
- Granovetter, M. (1973). The Strength of Weak Ties. *American Journal of Sociology*, 78(6), 1360–1380.
- Granovetter, M. (1985). Economic Action and Social Structures: The Problem of Embeddedness. *American Journal of Sociology*, 91(3), 481–510.
- Healy, K. (2011). Lucien Karpik Valuing the Unique: The Economics of Singularities. Princeton, Princeton University Press, 2010. *Socio-Economic Review*, 9(4), 787–791. <https://doi.org/10.1093/ser/mwr010> (Judgement and distinction).
- Hedström, P. (2005). *Dissecting the Social. On the Principles of Analytical Sociology*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511488801>
- Hennig-Thurau, T., Marchand, A., & Hiller, B. (2012). The relationship between reviewer judgments and motion picture success: re-analysis and extension. *Journal of Cultural Economics*, 36(3), 249–283. <https://doi.org/10.1007/s10824-012-9172-8>
- Hennig-Thurau, T., Wiertz, C., & Feldhaus, F. (2015). Does Twitter matter? The impact of microblogging word of mouth on consumers' adoption of new movies. *Journal of the Academy of Marketing Science*, 43(3), 375–394. <https://doi.org/10.1007/s11747-014-0388-3>
- Holbrook, M. B., & Addis, M. (2008). Art versus commerce in the movie industry: A Two-Path Model of Motion-Picture Success. *Journal of Cultural Economics*, 32(2), 87–107. <https://doi.org/10.1007/s10824-007-9059-2>
- Hutter, M. (2011a). Lucien Karpik Valuing the Unique: The Economics of Singularities. Princeton, Princeton University Press, 2010. *Socio-Economic Review*, 9(4), 791–794. <https://doi.org/10.1093/ser/mwr010> (Mapping a continent).
- Hutter, M. (2011b). Lucien Karpik: Valuing the unique. The economics of singularities. *Journal of Cultural Economics*, 35(4), 315–317. <https://doi.org/10.1007/s10824-011-9147-1>
- Ji, W., Keh, H. T., Singh, R., Sy-Changco, J. A., & Wang, X. (2015). Online movie ratings: A cross-cultural, emerging Asian markets perspective. *International Marketing Review*, 32(3/4), 366–388. <https://doi.org/10.1108/IMR-08-2013-0161>

- Kahneman, Daniel; Sibony, Olivier; Sunstein, Cass R. (2021): *Noise. A Flaw in Human Judgment*. New York: Little, Brown Spark.
- Karpik, L. (2007). *L'économie des singularités*. Gallimard.
- Karpik, L. (2010). *Valuing the unique: The economics of singularities*. Princeton Univ. Press.
- Karpik, L. (2011). *Mehr Wert: Die Ökonomie des Einzigartigen*. Campus.
- Katz, E., & Lazarsfeld, P. F. (1964). *Personal Influence: The Part played by People in the Flow of Mass Communications* (4th ed.). Free Press.
- Keuschnigg, M. (2015). Product success in cultural markets: The mediating role of familiarity, peers, and experts. *Poetics*, 51, 17–36. <https://doi.org/10.1016/j.poetic.2015.03.003>
- Keuschnigg, M., & Wolbring, T. (Eds.). (2015). *Experimente in den Sozialwissenschaften*. Nomos. <https://doi.org/10.5771/9783845260433>
- Kittel, B. (2015). Experimente in der Wirtschaftssoziologie: Ein Widerspruch? In M. Keuschnigg & T. Wolbring (Eds.), *Experimente in den Sozialwissenschaften* (pp. 79–104). Nomos. <https://doi.org/10.5771/9783845260433-82>
- Kraemer, K. (2017). Lucien Karpik: Mehr Wert. Die Ökonomie des Einzigartigen. In K. Kraemer & F. Brugger (Eds.), *Schlüsselwerke der Wirtschaftssoziologie* (pp. 507–514). VS Verlag für Sozialwissenschaften.
- Kumar, P., Divakaran, P., & Nørskov, S. (2016). Are online communities on par with experts in the evaluation of new movies? Evidence from the Fandango Community. *Information Technology & People*, 29(1), 120–145. <https://doi.org/10.1108/ITP-02-2014-0042>
- Lamont, M. (2012). Toward a Comparative Sociology of Valuation and Evaluation. *Annual Review of Sociology*, 38(1), 201–221. <https://doi.org/10.1146/annurev-soc-070308-120022>
- Lancaster, K. J. (1966). A New Approach to Consumer Theory. *Journal of Political Economy*, 74(2), 132–157.
- Levin, A. M., Levin, I. P., & Heath, E. C. (1997). Movie Stars and Authors As Brand Names: Measuring Brand Equity in Experiential Products. *Advances in Consumer Research*, 24, 175–181.
- Lieberman, E. (2006). Hollywood Economics: How Extreme Uncertainty Shapes the Movie Industry. *Movie Quarterly*, 59(3), 74–75. <https://doi.org/10.1525/fq.2006.59.3.74> (Book Review).
- Liu, A., Liu, Y., & Mazumdar, T. (2014). Star power in the eye of the beholder: A study of the influence of stars in the movie industry. *Marketing Letters*, 25(4), 385–396. <https://doi.org/10.1007/s11002-013-9258-x>
- Liu, H. (2016). A Structural Model of Advertising Signaling and Social Learning: The Case of the Motion Picture Industry. http://economics.usf.edu/PDF/ads_movie_HaiyanLiu_020916.pdf
- Liu, Y. (2006). Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue. *Journal of Marketing*, 70, 74–89.
- Maurer, A. (2014). Lucien Karpik, MehrWert. Die Ökonomie des Einzigartigen. *Soziologische Revue - Besprechungen Neuer Literatur*, 37(2), 213–215.
- Maurer, A., & Schmidt, C. (2019). Unsicherheit in der Wirtschaftssoziologie. Der Beitrag wirtschaftssoziologischer Ansätze zur Bearbeitung von Unsicherheit und Risiko. In H. Pelizäus & L. Nieder (Eds.), *Das Risiko – Gedanken übers und ins Ungewisse* (pp. 127–140). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-27341-5_5

- Mellet, K., Beauvisage, T., Beuscart, J.-S., & Trespeuch, M. (2014). A “Democratization” of Markets? Online Consumer Reviews in the Restaurant Industry. *Valuation Studies*, 2(1), 5–41. <https://doi.org/10.3384/vs.2001-5992.14215>
- Merton, R. K. (1968). The Matthew Effect in Science: The reward and communication systems of science are considered. *Science*, 159(3810), 56–63. <http://garfield.library.upenn.edu/merton/matthew1.pdf>
- Moon, S., Bergey, P. K., & Iacobucci, D. (2010). Dynamic Effects Among Movie Ratings, Movie Revenues, and Viewer Satisfaction. *Journal of Marketing*, 74(1), 108–121. <https://doi.org/10.1509/jmkg.74.1.108>
- Nelson, P. (1970). Information and Consumer Behavior. *Journal of Political Economy*, 78(2), 311–329.
- Nelson, R. A., Donihue, Waldman, D. M., & Wheaton, C. (2001). What’s an Oscar worth? *Economic Inquiry*, 39(1), 1–6. <https://doi.org/10.1111/j.1465-7295.2001.tb00046.x>
- Podolny, J. M [Joel Marc]. (2008). *Status signals: A sociological study of market competition*. Princeton University Press.
- Preda, A. (2007). The Sociological Approach to Financial Markets. *Journal of Economic Surveys*, 21(3), 506–533. <https://doi.org/10.1111/j.1467-6419.2007.00512.x>
- Ravid, S. A. (1999). Informations, Blockbusters, and Stars: A Study of the Movie Industry. *Journal of Business*, 72(4), 463–492.
- Ravid, S. A., Wald, J. K., & Basuroy, S. (2006). Distributors and movie critics: does it take two to Tango? *Journal of Cultural Economics*, 30(3), 201–218. <https://doi.org/10.1007/s10824-006-9019-2>
- Rutherford, A. (2001). *Introducing Anova and Ancova: A GLM approach. Introducing statistical methods*. Sage Publ.
- Sauer, C., Auspurg, K., & Hinz, T. (2020). Designing Multi-Factorial Survey Experiments: Effects of Presentation Style (Text or Table), Answering Scales, and Vignette Order. *Methods, Data, Analyses*, 14(2), 195–214. <https://doi.org/10.12758/MDA.2020.06>
- Schenk, P. (2021). Karpik in the Bottle: Can Judgment Devices Explain the Demand for Fine Wine? In: *Köln Z Soziol*. DOI: 10.1007/s11577-021-00794-4.
- Schoppek, W. (2015). Mehrebenenanalyse oder Varianzanalyse? *Zeitschrift Für Entwicklungspsychologie und Pädagogische Psychologie*, 47(4), 199–209. <https://doi.org/10.1026/0049-8637/a000136>
- Scott, A. J. (2005). *On Hollywood: The Place, the Industry*. Princeton University Press.
- Simonton, D. K. (2004). Movie Awards as Indicators of Cinematic Creativity and Achievement: A Quantitative Comparison of the Oscars and Six Alternatives. *Creativity Research Journal*, 16(2-3), 163–172. <https://doi.org/10.1080/10400419.2004.9651450>
- Situmeang, F. B.I., Leenders, M. A.A.M., & Wijnberg, N. M. (2014). The good, the bad and the variable: How evaluations of past editions influence the success of sequels. *European Journal of Marketing*, 48(7/8), 1466–1486. <https://doi.org/10.1108/EJM-08-2012-0493>
- Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Sage.
- Steiner, P. M., & Atzmüller, C. (2006). Experimentelle Vignettendesigns in faktoriellen Surveys. *Kölner Zeitschrift Für Soziologie Und Sozialpsychologie*, 58(1), 117–146. <https://doi.org/10.1007/s11575-006-0006-9>

-
- Tsao, W.-C. (2014). Which type of online review is more persuasive? The influence of consumer reviews and critic ratings on moviegoers. *Electronic Commerce Research*, 14(4), 559–583. <https://doi.org/10.1007/s10660-014-9160-5>
- Watts, D. J., & Salganik, M. J. (2011). Social Influence: The Puzzling Nature of Success in Cultural Markets. In P. Bearman, P. Hedström, P. Bearman, P. Hedström, D. J. Watts, & M. J. Salganik (Eds.), *The Oxford Handbook of Analytical Sociology*. Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199215362.013.14>
- Wolbring, T. (2017). Experimente in der Wirtschaftssoziologie. In A. Maurer (Ed.), *Handbuch der Wirtschaftssoziologie* (2nd ed., pp. 501–520). VS Verlag für Sozialwissenschaften.
- Zhuang, W., Babin, B., Xiao, Q., & Paun, M. (2014). The influence of movie's quality on its performance: Evidence based on Oscar Awards. *Managing Service Quality: An International Journal*, 24(2), 122–138. <https://doi.org/10.1108/MSQ-11-2012-0162>
- Zuckerman, E. W., & Kim, T.-Y. (2003). The critical trade-off: identity assignment and box-office success in the feature movie industry. *Industrial and Corporate Change*, 12(1), 27–67. <https://doi.org/10.1093/icc/12.1.27>

Appendix

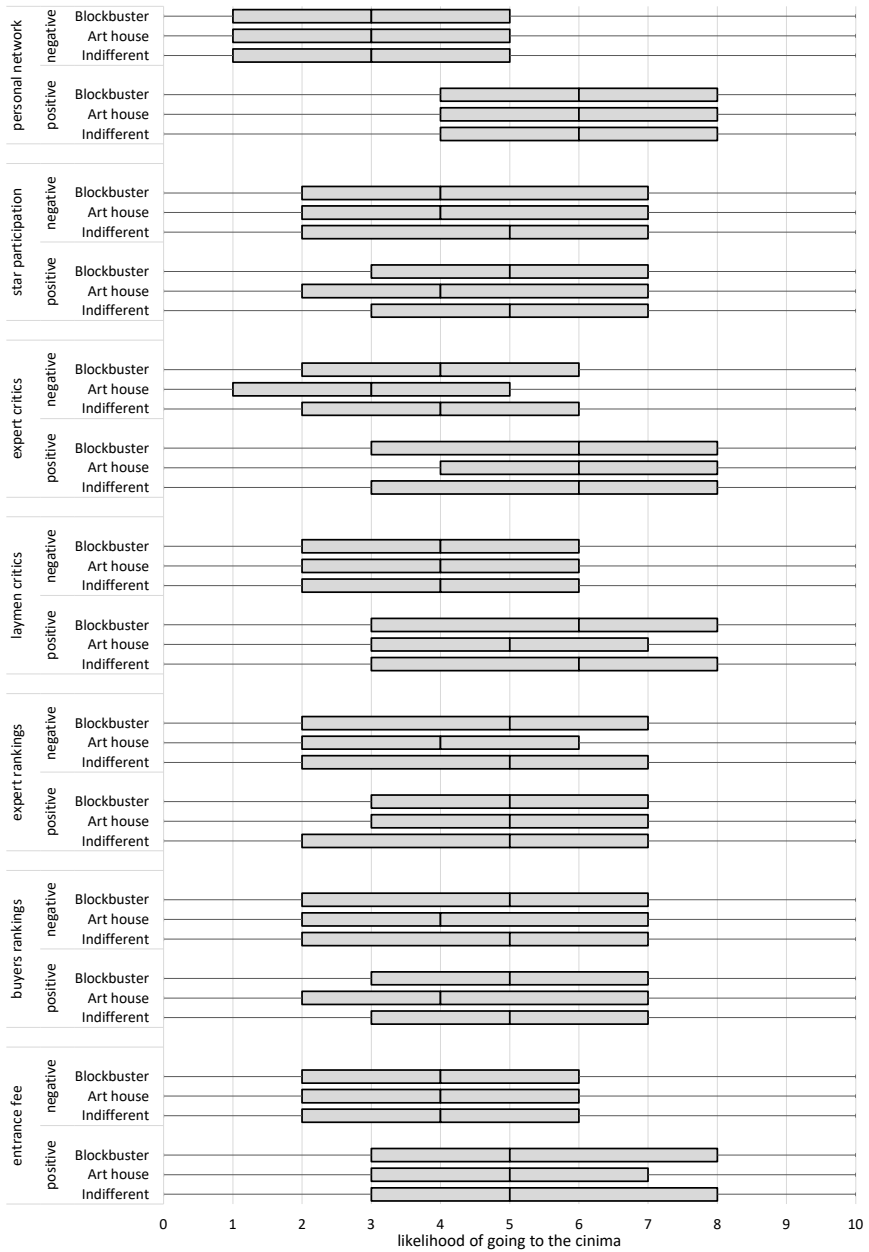


Figure 4 Boxplots for the likelihood of going to the cinema of the positive and negative levels, differentiated by movie preferences and dimensions

Fear of the Dark? A Systematic Comparison of Written Vignettes and Photo Vignettes in a Factorial Survey Experiment on Fear of Crime

Stefanie Eifler¹ & Knut Petzold²

¹ *Catholic University of Eichstätt-Ingolstadt*

² *Hochschule Zittau/Görlitz – University of Applied Sciences*

Abstract

Measuring attitudes with vignettes is frequently based on the assumption that the presented context information facilitates a better imagination of topics under study, serving for more valid responses as compared to more usual questionnaire methods. In this study, we focus on the presentation format of vignettes and assume that, in particular, the presentation of photo vignettes facilitates a close context approximation, hereby taking fear of crime from the perspective of broken windows theory as an example of use. A split ballot experiment within the framework of a cross-sectional online survey introduced a variation of the presentation format of a factorial survey experiment and allowed for measuring the difference between using either written vignettes or photo vignettes. While the split ballot experiment used a between-subjects design, each factorial survey experiment used a within-subjects design. The reported level of feelings of unsafety serves as a measure of fear of crime. Results show that, first, all dimensions of the factorial surveys predicted the respective level of fear of crime in both presentation formats, in the direction expected by broken windows theory. Measurement error seems slightly reduced within written vignettes. Second, presentation format-specific differences were observed for dimensions representing physical features of the setting, such as darkness, only, thereby slightly favouring photo vignettes. We finally discuss methodological implications of these results.

Keywords: Factorial Survey Experiments, Presentation format, Written vignettes, Photo vignettes, Broken Windows Theory, Fear of Crime



Our study focuses on factorial survey experiments that are used to measure normative judgements, subjective beliefs or behavioural intentions (cf. Beck & Opp, 2001; Jasso, 2006) through respondent's answers to a number of brief descriptions of hypothetical situations, persons or objects called vignettes (Auspurg & Hinz, 2015; Jasso, 2006; Rossi, 1979; Rossi & Anderson, 1982). Due to their supposed advantages, vignettes have been increasingly applied in surveys (cf. Auspurg & Hinz, 2015; Liebig et al., 2015; Mutz, 2011; Wallander, 2009). First, because of the systematic variation of several features or dimensions, the relative weight of these dimensions with regard to the responses can be determined. Second, effects of a self-selection driven by the respondents' interests can be neutralised through randomisation. Variation and randomisation are also features of a random experiment; thus, third, vignette analyses allow for a causal interpretation of the effects of situational features or vignette dimensions. It is usually stated that, fourth, vignettes comprise more detailed and more concrete information on the phenomena meant and, therefore, facilitate a more standardised imagination of the situation across respondents and less use of general heuristic principles by respondents, hence inducing them to report their *true* opinions (e.g., Shamon et al., 2019).

One main argument for using vignettes is that the presentation of information on the situational context helps to achieve a close proximity to the reality of everyday life. Accordingly, several authors have pointed out that vignettes allow to mirror situations of everyday experience and, thus, to bring individual answers in line with real-life judgement formation or decision-making (cf. Alexander & Becker, 1978; Armacost et al., 1991; Finch, 1987). However, the presentation form of vignettes shapes the results of vignette-based measurements, for example, a detailed or sparse presentation (e.g., Eifler & Petzold, 2014) or the presentation of vignettes in running text or tabular format (Sauer et al., 2020; Shamon et al., 2019). So far, it is still an open question whether factorial survey experiments actually help to improve measurement quality of normative judgments, subjective beliefs or behavioural intentions.

In principle, there are different formats of presenting vignettes within the framework of a survey: the situation can either be described in a written form or presented by visual stimuli, for example, by videos, photos or pictures. While the majority of studies apply written vignettes (Wallander, 2009), some studies use solely video vignettes (Krysan et al., 2009) or solely photo vignettes (Golden III

Acknowledgements

The authors thank two anonymous reviewers for helpful comments to a previous version of this manuscript. Both authors contributed equally to this work.

Direct correspondence to

Stefanie Eifler, Catholic University of Eichstätt-Ingolstadt, Ostenstraße 26,
85072 Eichstätt
E-mail: stefanie.eifler@ku.de

et al., 2001). Another study combines different written and photo information in vignettes (Havekes et al., 2013). Beyond scarce applications, only two studies compared presentation format differences systematically (Eifler, 2007; Rashotte, 2003). Both authors found systematic differences between verbal and visual presentation formats, but they also stated that much more research is required to clearly determine differing results for various types of stimuli.

Against this background, our study is particularly devoted to the presentation format of vignettes. It is a largely open question whether or not written and photo vignettes lead to corresponding or diverging responses and whether or not the effects of situational dimensions in a factorial survey experiment depend upon the presentation format used.

To fill this research gap, we start from psychological approaches which state that different processes of recognition, information processing and remembering verbal and visual information apply. In particular, we apply the *Dual Coding Theory* (DCT) suggested by Paivio (1979) and Sadoski and Paivio (2013) to the systematic analysis of presentation format differences concerning the use of written vignettes or photo vignettes. Taking the example of the broken windows theory, which we employ for the prediction of fear of crime (Keuschnigg & Wolbring, 2015; Keizer et al., 2014; Kelling & Coles, 1996; Wilson & Kelling, 1982), we use vignettes that describe or visualise varying situations of everyday experience. We assign respondents randomly to one of the two presentation formats. By doing so, we demonstrate both presentation format correspondence and presentation format differences.

In the next section, we present a model concerning the role of the presentation format of vignettes and derive testable hypotheses. We analyse the assumptions empirically on the basis of a split ballot experiment among the population of students from a German university. Finally, we discuss our findings critically and consider methodical implications. Overall, our study demonstrates both validity aspects of factorial survey experiments using different presentation formats of vignettes and theoretically predictable differences between these presentation formats.

A Systematic Comparison of Written Vignettes and Photo Vignettes

Just as in the case of answering survey items, we can use the general model of the response process in surveys by Tourangeau (1984) and Tourangeau et al. (2000), in order to delineate the process of responding to vignettes. According to this model, a respondent who is asked a question first has to interpret the question's content (interpretation), subsequently has to retrieve information from the memory

(retrieval), form an opinion (judgement) and then bring the answer into line with the predefined response format (response selection). Transferred to the measurement with vignettes, a subject first has to interpret the situation and the question presented and has to retrieve information from his/her memory referring to it, before he/she can form an opinion and provide an answer.

With regard to factorial survey experiments, several authors have emphasised the idea that using vignettes facilitates a standardised presentation of information about the situations under study (Auspurg & Hinz, 2015; Mutz, 2011; Jasso, 2006; Rossi, 1979; Rossi & Anderson, 1982). In the eyes of Shamon et al. (2019), this leads to a more unified retrieval – called “information intake” (p. 4) by the authors – of relevant information from the memory across subjects. Accordingly, the retrieval stage of the response process is assumed to be characterised by a higher level of interindividual comparability in factorial survey experiments as opposed to survey items.

While the majority of studies apply written vignettes (Wallander, 2009), several researchers have suggested to use visual stimuli within the framework of factorial survey experiments because visual stimuli like video clips, photos or pictures allow for a more natural representation of the situations under study, indicating a clear preference for video vignettes (Caro et al., 2012a, 2012b; Dinora et al., 2020; Golden III et al., 2001; Goyal et al., 2017; Havekes et al., 2013; Hughes & Huby, 2004; Krysan et al., 2009; O'Donnell et al., 2007; Rashotte, 2003).

To our knowledge, only two studies compared observed responses to both written and visual stimuli (Eifler, 2007; Rashotte, 2003). Rashotte (2003) examines what information people receive and use in forming effective responses when observing written versus visual stimuli on social events. In her study, readers of written descriptions of events and viewers of videotapes use different pieces of information in forming impressions based on stimuli type (Rashotte, 2003). While visual cues of nonverbal behaviours appear clearer in videotapes and viewers need less information than readers to get an impression of it, viewers use the same information as readers to evaluate object-persons themselves. The results are consistent with the idea that visual stimuli provide more information and allow for a richer picture of social events. The assumption that visual presentations provide more accurate representations of situations and, thus, evoke more valid responses is also tested by Eifler (2007). Behavioural observations and vignette analyses with visual and verbal material were carried out with regard to three forms of deviant behaviour in everyday life, showing that frequencies of (intended) deviant behaviour were related to the presentation formats. Written vignettes lead to an overestimation of the frequencies of crossing a red traffic light and to an underestimation of the frequencies of cycling through a red traffic light. While deviant behaviour to ignore a ‘lost letter’ is overestimated by all respondents, the degree of overestima-

tion is smaller in the face of a visual vignette. In both studies, it becomes clear that the role of the presentation format needs more clarification.

Presentation Format and Information Processing

So far, little is known about potential differences between written vignettes and vignettes presenting pictures or photos. In particular, there are – to our knowledge – no systematic theory-guided approaches that would help to explain *why* visual stimuli should be superior to the usual verbal presentations of vignettes. Therefore, we will introduce theoretical ideas from cognitive psychology in order to explain format differences in factorial survey experiments.

We, thereby, start from psychological approaches which state that different processes of recognition, information processing and remembering verbal and visual information apply. In particular, we refer to the DCT suggested by Paivio (1979) and Sadoski and Paivio (2013), which posits the idea that verbal and visual information is coded differently in the human brain.

This approach starts from the idea that there are two coding systems in human memory: one responsible for language or verbal information and the other responsible for pictures or non-verbal information: “In DCT, the linguistic coding system is referred to simply as the verbal code or system, and the nonverbal coding system is often referred to as the imagery code or system because its main functions include the analysis of external scenes and the generation of internal mental images” (Sadoski & Paivio, 2013, p. 29). Both systems overlap and can operate simultaneously in principle. Processing verbal and/or visual information generates “internal mental images” (Sadoski & Paivio, 2013, p. 29) which represent information about situations. It is assumed that mental images of situations match experiences with the same situations (Kosslyn & Pomerantz, 1977; Kosslyn, 1981). Concerning the prediction of systematic presentation format differences, this central assumption would require a specification of particular features of a hypothetical situation with regard to using either written or photo vignettes.

What is crucial with regard to these mental images is that written and visual information about situations is processed by both systems but in a different way: Written information is processed *sequentially* (i.e., by the verbal coding system first and by the non-verbal coding system subsequently), and visual information is processed *simultaneously* by both coding systems at a time (Paivio, 1979; Sadoski & Paivio, 2013). Because of the sequential processing of verbal information, written vignettes can elicit diverging encoding processes by readers, thus leading to diverging visualisations in memory between subjects. In studies on learning and memory, the thesis that verbal information is visualised by readers was supported (Kosslyn, 1981). Because of the simultaneous processing of visual information, photo vignettes facilitate a standardised perception of the concrete situation without any

loop way, thus leading to corresponding mental images of the presented situations between subjects. Correspondingly, Hanna and Loftus (1993) pointed at qualitative differences between verbal and visual information processing. In addition, Harper (2002) emphasised that the parts of the brain that process visual information are evolutionarily older than the parts that process verbal information; thus, images might evoke deeper elements of human consciousness than do words.

While there is much research activity concerning functions of visual memory, like remembering or recalling natural scenes or – more generally – everyday experience (Brockmole, 2009; Findlay & Gilchrist, 2003; Luck & Hollingworth, 2008), there are not more than a handful of studies that are devoted to a systematic comparison of the cognitive processes involved in remembering and recalling both verbal and visual information. Overall, neurophysiological studies have shown that visual stimuli are remembered and recalled more easily than verbal stimuli (Bower, 1970; Shepard, 1967). Correspondingly, a systematic comparison between visual and verbal information revealed that photos are remembered better than words, which was explained in the following way: “(...) pictures contain distinctive cues which make them more discriminable than their labels and this discriminability enhances memory for pictures compared to their labels” (Jenkins et al., 1967, p. 306). McCloud (1994) summarised these differences and stated that verbal information is *perceived*, while visual information is *received*.

From the theoretical considerations presented so far, we conclude that differences in processing verbal and visual information exist. Verbal information requires more extensive information processing by a reader and more background knowledge, whereas visual information presents the information directly. Following this train of thought, photos can be considered to mirror real life (Manghani, 2013, Rose, 2012). According to Barthes (1977, p. 17), a photo is “(...) not the reality but at least it is its perfect *analogon* and it is exactly this analogical perfection which, to common sense, defines the photograph (...): *it is a message without a code*”. Accordingly, photos are a concrete point of reference for all who are confronted with them (Collier Jr., 1957; Collier & Collier, 1986). In a similar way, other authors highlight the advantages of presenting photos: “Showing many things at once is a tremendous strength that reflects the all-at-once nature of lived experiences – a reality that is often impossible to communicate through linear textual narratives” (Marion & Crowder, 2013, p. 31).

Therefore, with regard to factorial survey experiments, photo vignettes not only allow for a more realistic presentation of the situations under study but also for evoking the feeling of experiencing the particular situation. While written vignettes facilitate a sequential presentation of information in the form of short stories, photo vignettes present the information simultaneously in the form of pictures, thereby activating visual and verbal mental representations and leading to emotional arousal at the same time.

Explanatory Model and Hypothesis

It follows from the above explicated theoretical ideas, in particular from DCT (Paivio, 1979; Sadoski & Paivio, 2013), that both presentation formats, written vignettes and photo vignettes, lead to mental images that include a visualisation of the presented situation. Therefore, we would expect mostly corresponding results between both presentation formats in a factorial survey experiment with regard to the direction of effects of situational dimensions. In principle, both written vignettes and photo vignettes should facilitate a representation of the same higher order constructs. Nevertheless, as for the simultaneous information processing of visual information, we would expect advantages of using photo vignettes with regard to the strength of effects of situational dimensions.

In order to test these assumptions, we took the broken windows theory (Kelling & Coles, 1996; Keizer et al., 2014; Keuschnigg & Wolbring, 2015; Lewis & Salem, 1986; Skogan, 1990; Wilson & Kelling, 1982) as an example of use. Amongst other topics, this approach has been applied to the analysis of fear of crime. The theory – also referred to as the *Disorder Model* of fear of crime – specifies features which are assumed to be perceived as cues of normative compliance in urban neighbourhoods. Because these features can be both described and pictured, the approach seems particularly suited for a systematic comparison of written vignettes and photo vignettes within the framework of a factorial survey experiment. In addition, visual methods have been used in the analysis of fear of crime because of their feasibility for presenting the context of crime-related cognitions and emotions (Vanderveen, 2018). We tie in with this tradition in principle and extend it to the systematic comparison of presentation format differences in factorial survey experiments.

Within the framework of the disorder model, one refers to the features of urban neighbourhoods that are called “signs of incivility” (Hunter, 1978). These are signs of non-compliance with behavioural norms like littering, graffiti on facades, destruction and decay of buildings or unsupervised youth (Hunter, 1978). In particular, physical signs of disorder, like plaster crumbling of the wall, are distinguished from social signs of disorder, like teenagers hanging around and drinking alcohol (Hunter, 1978; Skogan, 1978; Taylor, 1999). It is assumed that signs of incivility serve as cues for the likelihood of norm enforcement in specific situations. They indicate a failure of informal control processes and call forth perceived victimisation risks, which, in turn, are reflected in higher levels of fear of crime.

For our systematic comparison of written and photo vignettes, we took signs of physical and social disorder, and introduced them as dimensions into factorial survey experiments. In both formats, the situations presented in the vignettes were systematically varied regarding the same dimensions: observability of place, physical decay and littering, unsupervised youth, adult passers-by, video surveillance,

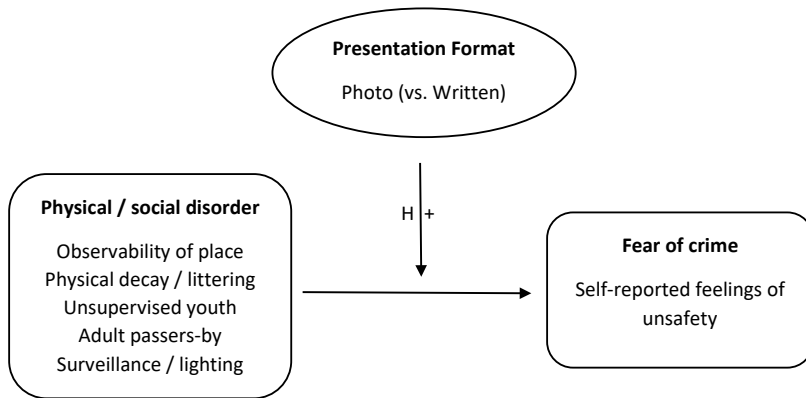


Figure 1 Underlying theoretical model

and lighting. If a setting exhibits physical and social features that indicate a high level of disorder, the level of fear of crime should be increased. We assume that, first, the direction of these influences will be comparable across the presentation formats of the factorial surveys. However, driven by different information processing, we assume stronger effects for the presentation in photo vignettes. Figure 1 shows the vignette dimensions and the hypothesis considered in the model. We suppose that the influences of physical and social features of the setting will be stronger in factorial surveys based on visual stimuli.

Hypothesis:

The effects of physical and social features of the setting that indicate a high level of disorder are stronger in a factorial survey employing photo vignettes compared to a factorial survey employing written vignettes.

Methods

Our empirical examination of presentation format differences in factorial surveys took place within the framework of a split ballot experiment (Benson, 1941) that was part of a survey on attitudes towards safety in public places. We conducted this survey at the Martin-Luther-University Halle-Wittenberg, Germany, in 2014 (Schwarzbach & Eifler, 2020).

Procedure, Data

We set up the study as a web survey using Lime Survey. The survey period was between 9 December 2013 and 17 January 2014. We invited respondents to participate in the survey by an email, providing them with a link to the online questionnaire. The survey followed the “Tailored Design Method” (TDM) (Dillman et al., 2009).

Sample

The survey included the full population of all enrolled students of the Martin-Luther-University Halle-Wittenberg, Germany. We administered the invitation to take part in the survey with the help of the registrar’s office, by sending an email to the full population of around $N = 20,000$ students. An overall number of $n = 1,149$ students completed the survey. Their mean age was 23.8 years, and 65.8% of them were female. Unfortunately, comparisons with the full population were not possible because no information about age and gender was provided for all enrolled students.

Operationalisation, Measurement

In the following, we describe the experimental design that we used to analyse the effects of using varying presentation formats in factorial surveys. We also describe the design of these factorial surveys. A complete project documentation including the questionnaires is available online for both transparency and replication purposes (Schwarzbach & Eifler, 2020).

Independent Variables

In our study, the subjects responded to a factorial survey using either written vignettes or photo vignettes (i.e., we used a between-subjects design for the split ballot experiment). We randomly assigned each subject to one of the two presentation formats.

To analyse the difference between the presentation formats, we used two factorial surveys based on the same 2^4 -within-subjects design. The factorial surveys referred to signed of social and physical disorder in urban neighbourhoods that had been used in previous studies (Piquero, 1999; Taylor, 1999). Table 1 illustrates the experimental design.

We pictured the *observability of the place* as either a wide square or a narrow pedestrian underpass. Facades covered with graffiti, empty beer bottles and other garbage around (high physical disorder) versus a clean and tidy environment (low physical disorder) represented *physical decay and littering*. We indicated

Table 1 Experimental Design

Dimensions	Levels		
	1	2	3
1 Observability of place	Wide square	Pedestrian underpass	
2 Physical decay, littering	No	Yes	
3 Unsupervised youth	Couple goes for a walk	Teenagers hanging around	
4 Adult passers-by	Passers	No passers	
5 Surveillance / lighting	Bright situation through lighting	Gloomy situation, but video surveillance	Gloomy situation without surveillance

Note: Cartesian product of dimensions and levels $2 \times 2 \times 2 \times 2 \times 3 = 48$ unique situations

unsupervised youth by teenagers hanging around (high social disorder) versus a young couple going for a walk (low social disorder). The presence (high social control) or absence (low social control) of *adult passers-by* referred to the respective dimension. *Surveillance and lighting* were part of one dimension including three levels: the presence of CCTV in a gloomy setting (video surveillance), a bright setting through the presence of street lighting (lighting) and the absence of video surveillance and street lighting in a gloomy setting (gloomy setting). We decided to use a dimension comprising three levels because a full combination of lighting and CCTV seemed inappropriate, as CCTV requires sufficient lighting. A group of experts ($n = 15$), composed of graduate students from the social sciences with a special training in the factorial survey approach, rated the correspondence between written vignettes and photo vignettes to facilitate a test of the presentation format.

From a full combination of the above explained dimensions and their levels, a universe of 48 vignettes was obtained. Given this large number of vignettes, we decided to present vignette sets to our subjects. Thereby, we assured to facilitate an estimation of all main effects of the vignette dimensions. Following the recommendations given by previous methodological studies on factorial surveys (for an overview, see Auspurg & Hinz, 2015), we used six sets of eight vignettes each. We presented one instruction to the respondents for both written vignettes and photo vignettes: "In the following, we ask you to judge a number of situations. We are interested in your feelings of safety or unsafety in these situations. Please put yourself in these situations:" Figure 2 shows an example of two written vignettes and their respective photo vignette counterparts.

Example 1: Photo vignette**Example 1: Written vignette**

You are on a wide square. The place is only dimly lit but you will see a sign saying “This area is under video surveillance”. The area looks neat and tidy. You realise two teenagers who hang around and drink alcohol. There are some additional adults nearby.

Example 2: Photo vignette**Example 2: Written vignette**

You are on a wide square. The place is brightly lit. The area looks neat and well kept. You see a young couple going for a walk. There are some additional adults nearby.

Figure 2 Examples of photo vignettes and written vignettes

Table 2 Quality of randomisation and variation, sample: estimation model

	Total		Written Vignettes		Photo Vignettes	
	N	Percent / M	N	Percent / M	N	Percent / M
<i>Vignette treatments</i>						
Observability of place						
Wide square	4046	49.99	1921	50.34	2143	49.69
Pedestrian underpass	4065	50.01	1895	49.66	2170	50.31
			$\chi^2 = 0.3460, p = 0.556$			
Physical decay, littering						
No	4091	50.38	1952	51.15	2139	49.59
Yes	4038	49.62	1864	48.85	2174	50.41
			$\chi^2 = 1.9679, p = 0.161$			
Unsupervised youth						
Couple goes for a walk	4096	50.39	1924	50.42	2172	50.36
Teenagers hanging around	4033	49.61	1892	49.58	2141	49.64
			$\chi^2 = 0.0029, p = 0.957$			
Adult passers-by						
No	4095	50.38	1919	50.29	2176	50.45
Yes	4034	49.62	1897	49.71	2137	49.55
			$\chi^2 = 0.0217, p = 0.883$			
Surveillance / lighting						
Bright situation, lighting	2727	33.55	1260	33.02	1467	34.01
Gloomy situation, video surv.	2719	33.45	1275	33.41	1444	33.48
Gloomy situation, no surv.	2638	33.01	1281	33.57	1402	33.51
			$\chi^2 = 1.2927, p = 0.524$			
<i>Questionnaire Characteristics</i>						
Vignette set						
1	1320	16.24	587	15.38	733	17.00
2	1636	20.13	732	19.18	904	20.96
3	1300	15.99	623	16.33	677	15.70
4	1356	16.68	651	17.06	705	16.35
5	1169	14.38	546	14.31	623	14.44
6	1348	16.58	677	17.74	671	15.56
			$\chi^2 = 1.6331, p = 0.897$			
Presentation format						
Written vignettes	3816	46.94				
Photo vignettes	4313	53.06				
<i>Total</i>						
N _{vignettes}	8129	100.0	3816	100.0	4313	100.0
N _{probands}	1019	100.0	479	100.0	540	100.0

Note: Test statistics for age, gender, deck at probands level.

Table 3 Parallelisation of experimental groups, sample: estimation model

	Total		Written Vignettes		Photo Vignettes	
	N	Percent / M	N	Percent / M	N	Percent / M
<i>Respondents' Characteristics</i>						
Gender						
Female	665	65.78	309	64.92	356	66.54
Male	346	34.22	167	35.08	179	33.46
			$\chi^2 = 0.2959, p = 0.586$			
Age		23.82		23.90		23.75
			$t = 0.5715, p = 0.568$			
Partner						
Yes	571	57.56	286	59.71	285	52.88
No	421	42.44	180	37.58	241	44.71
			$\chi^2 = 5.2303, p = 0.022$			
$N_{\text{respondents}}$	1019	100.0	479	100.0	540	100.0

Note: Test statistics for age, gender, deck at respondents' level.

Each respondent answered eight vignettes, which resulted in a full estimation sample of $n = 8,129$ judged vignettes. As for the presentation formats under study, the full estimation sample included $n = 3,816$ for the written vignettes and $n = 4,313$ for the photo vignettes. To assess the design's accessibility to systematic group comparisons, we evaluated the randomisation of subjects across the experimental conditions of the split ballot experiment for analysing presentation format differences of the factorial surveys. To do so, we considered whether a parallelisation with regard to the split ballot experiment emerged on the basis of the full estimation model (Table 2).

As depicted from Table 2, there are no substantial differences with regard to the distribution of the vignette dimensions across the two levels of the split ballot experiments (i.e., the presentation of the factorial surveys either using written vignettes or photo vignettes). Subsequently, we examined the vignette dimensions and the respondents' characteristics – age and gender – for a uniform distribution across both vignette presentation modes and show the results in Table 3, which reveals that randomisation of subjects to the presentation formats led to mostly parallel groups with regard to respondents' gender, age and partnership status.

Dependent Variable

The key dependent variable referred to the level of fear of crime when facing the situations described by means of the vignettes. To measure this, we used the stan-

Table 4 Distributions of the dependent variable “Fear of Crime” as reported feelings of unsafety; sample: estimation model

Feelings of unsafety	Both		Written Vignettes		Photo Vignettes	
	N	Percent	N	Percent	N	Percent
Very safe	1880	23.13	876	22.96	1004	23.28
Safe	3619	44.52	1715	44.94	1904	44.15
Unsafe	2028	24.95	971	25.45	1057	24.51
Very unsafe	602	7.41	254	6.66	348	8.07
Total	8129	100.00	3816	100.00	4313	100.00
M	1.166		1.158		1.174	
SD	0.866		0.852		0.878	

T-test: $t = -0.813$, $p = 0.208$

U-test: $z = -0.471$, $p = 0.638$

dard indicator for fear of crime (i.e., the level of feelings of unsafety). Immediately after presenting each vignette, we asked the following question: “How safe would you feel in this situation?”. In response to the question, the subjects used a rating scale (0: very safe; 1: safe; 2: unsafe; 3: very unsafe) for shaping their answer. Table 4 shows the resulting distributions for the full estimation sample.

It follows from Table 4 that, independent of the respective presentation format, most subjects reported a lower level of feelings of safety in public places. A comparison between formats revealed no relevant differences between presentation formats. This reflects that both formats stimulated similar responses on the aggregate level (i.e., across all vignettes).

Method of Analysis

All subsequent analyses refer only to those subjects who considered the factorial surveys as realistic. In conjunction with an evaluation of the online-questionnaire in both formats, we asked the subjects to indicate whether they could imagine themselves in the situation that is presented in the vignettes using a dichotomous response format (0: no; 1: yes). It can be taken from Table A-1 in the appendix that by far, the majority of all subjects evaluated the situations presented in both the written vignettes and the photo vignettes as realistic ($n = 1,019$), while only a minority did not ($n = 188$). If the latter group is included in the estimation, only slight differences between all subjects and those who recognised the vignettes as realistic emerged (see Table A-1 in the appendix). For reasons of accurateness, we

decided to use the sample of respondents who evaluated vignettes as being realistic only.

We measured our outcome variable, the level of self-reported feelings of unsafety as an indicator of the level of fear of crime, on a rating scale with four stages, which we interpreted as quasi-metric so that regression models for continuous dependent variables can be applied.

As each respondent assessed a number of vignettes describing varying situations of physical and social disorder, the data structure is hierarchical (Hox et al., 1991; Jasso, 2006). To consider the multi-level structure, we used random intercept fixed slope models, which account for the variation in the outcome variable between respondents (e.g., Snijders & Bosker, 2012). Due to the rather small number of observations at the first level, which is a consequence of the restricted size of the vignette sets, we estimated only the intercept with a random component.

Our primary interest lies in a direct comparison of the factorial surveys across both presentation formats. Accordingly, we estimated a joint model and included multiplicative cross-level-interaction terms between the presentation format at level 2 and all treatment dimensions at level 1. This allows for the estimation of the format's main effect and effects of the vignette dimensions and their levels conditional to the presentation of written or photo vignettes. Accordingly, our estimation strategy can be noted as follows:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \beta_2 f_j + \beta_3 f_j X_{ij} + \nu_j + \varepsilon_{ij} \quad ; i = 1, \dots, n; j = 1, \dots, m \quad (\text{Eq. 1})$$

Y_{ij} : Reported feelings of unsafety of a respondent j towards a vignette i

X_{ij} : Vector of disorder characteristics varied in vignettes

f_j : Format of presentation to each respondent (photo or written)

ν_j : Error term at respondent level

ε_{ij} : Error term at vignette level

In addition to the analytic strategy presented, the results have undergone a number of checks to prove for their robustness. In a first step, we checked for successful randomisation of the vignettes across respondents, by comparing the dimensions' main effects in models with and without control variables at the respondents' level (see Table A-2 in the appendix). We considered respondents' gender, age and relationship status as sociodemographic covariates, completed by their stated feelings of fatigue when answering the vignettes. All coefficients regarding effects of the vignette dimensions are quite similar between both models, reflecting that randomisation resulted in balanced covariates at the respondents' level. We further compared the random effects model with a fixed effects model using the Hausman test, which revealed only slight and non-substantial differences in coefficients. The

checks indicate that our results are remarkably robust what further confirms that randomisation worked well at both stages of our experimental design. On the basis of these finding, we estimated the presented model without covariates in order to ensure for less missing data due to non-response. As can further be taken from Table 2, the six vignette sets were not assigned to the respondents with exactly the same number during data collection. To account for systematic differences in judgements between vignette sets, we have fixed the effects of the sets in all regression models.

Furthermore, the outcome measurement at a four-point response scale may violate the requirements for linear modelling. Therefore, we have replicated the main effects model in both an ordinal logit model (Table A-3 in the appendix) and a binary logit model (Table A-4 in the appendix). For the latter, we dichotomised the outcome variable using a median split. Although the absolute values of the coefficients cannot be compared directly due to different modelling and scaling, they nevertheless show the same directions and relative strengths within the models. Therefore, the results indicate the robustness of our results. In addition, a comparison between a simple linear regression that neglects the nested data structure and the multilevel model also corroborates our interpretation (Table A-5 in the appendix).¹ We report p-values and confidence intervals to facilitate interpretation.²

-
- 1 Both, data and codes concerning the analyses strategies will be provided by the authors for replication purposes upon reasonable request.
 - 2 Applying conventional methods of statistical inference is justified even though we did not draw a random sample of respondents for two main reasons: At first, random assignment in experiments reflects data generation through known probability procedures which facilitates formally capturing uncertainties (cf. Berk et al., 1995). Randomisation of subjects to treatments allows for attributing differences between treatments to randomisation error. This justifies testing null hypotheses for treatment effects although statistical inferences apply only to the respondent sample actually used (Edgington, 1966). At second, we invited the total population of students of a German university and consider this population as a realisation from some super population, i.e., a target population which is wider than the actual population under study (Alexander, 2015). On the background of these considerations, we consider our sample population as an equivalent of a random sample which may be analysed on the basis of frequentist methods (Berk et al., 1995).

Results

In this section, we present the results of a random intercept multilevel regression model with interaction terms between the presentation format and all vignette dimensions. The effects of all vignette dimensions conditional on the presentation format are plotted in Figure 3.³

As already described above, the mean values of the response scales do not differ across the presentation formats. This finding is reflected by the very small and insignificant coefficient of the main effect of the presentation format in the regression model again. This suggests that both written and photo vignettes generate similar response patterns, at least on the aggregate level including all vignettes. This may be interpreted as a sign of a basic level construct validity in both presentation formats.

It follows from Figure 3 that the effects of the vignette dimensions support the broken windows theory for both presentation formats. The self-reported feeling of unsafety is the stronger the more signs of physical or social disorder are present in a setting as compared to the respective reference categories. The respondents feel more unsafe if a pedestrian underpass is shown instead of a wide square and if there is physical decay and littering indicated by graffiti and garbage lying around instead of a clean and tidy setting. They also report more concern about teenagers hanging around than by a couple walking. Adult passers-by reduce their feelings of unsafety. Compared to a bright scenery resulting from lighting, the respondents feel less safe both in a gloomy situation and when there is video surveillance. The strongest effect is revealed for teenagers hanging around, whereas not being able to overlook a place shows the least effect.

In our hypothesis, we stated that the effects of features of the setting that indicate a high level of physical or social disorder upon the level of fear of crime are stronger in a factorial survey employing photo vignettes compared to a factorial survey employing written vignettes. The conditional effects reveal that there are clear differences in the effects of the vignette dimensions between the two presentation formats, particularly with regard to the dimension of surveillance and lighting. Compared to bright lighting, both video surveillance and a gloomy scenario increase the level of self-reported feelings of unsafety in the photo vignettes significantly stronger than in the written vignettes. Put differently, this means that the level 'bright situation through lighting', compared to the levels 'gloomy situation without surveillance' and 'gloomy situation, but video surveillance', reduces feelings of unsafety much stronger in the photo vignettes than in the written vignettes. Accordingly, the statistically significant coefficients of the interaction terms reveal

3 For reasons of an easier interpretation, the most important results are shown as graphics. The information on the complete regression model can be found in Table A-6 in the appendix.

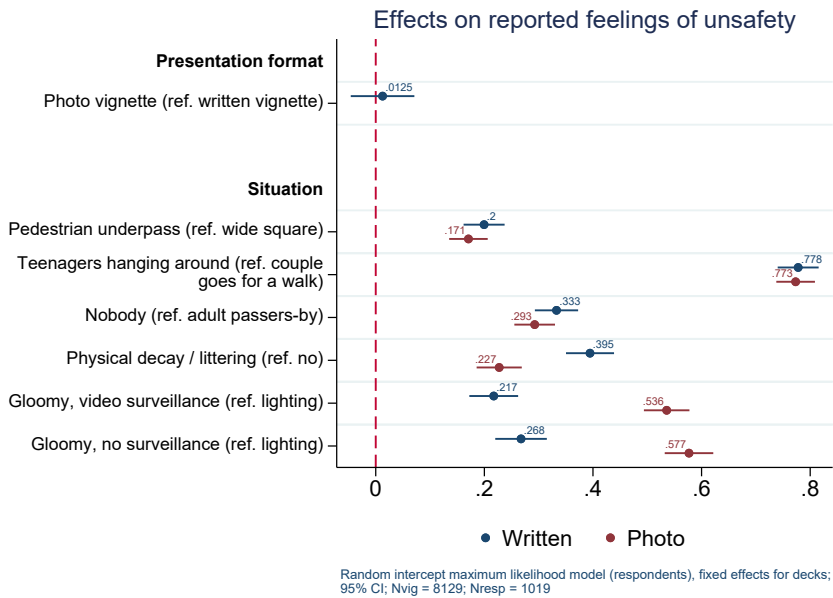


Figure 3 Results of interaction effects model on reported feelings of unsafety (effects conditional to presentation format)

remarkable differences between the two presentation formats both for video surveillance ($b = 0.318, p = 0.000$) and for a gloomy situation without surveillance ($b = 0.309, p = 0.000$) by about a third unit at the rating scale. These findings do fully support our hypothesis.

In addition, a significant interaction effect between the level of physical decay/littering and the presentation format is revealed ($b = -0.167, p = 0.000$). However, the direction of this effect is contrary to our assumption, since physical decay and littering, compared to a clean and tidy environment, increases the self-reported feelings of unsafety more in the written vignettes than in the photo vignettes. This finding, therefore, contradicts our hypothesis.

The remaining signs of disorder of teenagers hanging around instead of a couple walking, the absence of adult passers-by instead of their presence and a narrow pedestrian underpass instead of a wide square show very similar regression weights that do not differ significantly. This indicates that these stimuli evoked comparable response behaviour in both presentation formats.

In summary, we found evidence that the respondents reacted more strongly to the dimension of surveillance/lighting and less to the level of physical decay/littering in the photo vignettes. It can be stated that differences between the presentation formats primarily occurred with regard to signs of physical disorder but hardly

with regard to signs of social disorder. Regarding the latter, written vignettes and photo vignettes seemed to work similarly in both presentation formats.

In addition to differences between treatment effects, we examined possible differences in measurement error by analysing the residuals in separate multi-level models for each presentation format (see Table A-7 in the appendix). Both the log likelihood and information criterion (AIC) show that the model for written vignettes fits data slightly better than the model for photo vignettes. This may suggest that photo vignettes result in data of poorer quality, while written vignettes gain better data. However, a more differentiated comparison of the variance of self-reported feelings of unsafety within and between subjects shows that the poorer model fit is solely due to the variance between respondents, while the variance within respondents is more or less equal for both presentation formats. Accordingly, R^2 coefficients for the variation within subjects hardly differ, while R^2 for the variation between subjects indicates a better fit for the model concerning written vignettes. While the responses differ somewhat more between the respondents in the photo format, the consistency of the responses is roughly the same for both presentation formats. That is, measurement invariance within a respondent is the same regardless of the presentation format, indicating that one of the presentation formats does not force them to give worse answers.

Furthermore, we already assessed the subjective costs of the administration of each format elsewhere (Eifler et al., 2021). These former analyses revealed that dropout rates do not differ between presentation formats, while processing time and self-reported fatigue are reduced when administering a questionnaire including photo vignettes. Concerning dropout rates, results do not indicate that factorial surveys based on photo vignettes would be superior with regard to respondent's willingness to participate. Hence, while the quality of individual responses is largely equal for both presentation formats, evaluating photo vignettes goes along with reduced administration costs in terms of processing time and subjective cognitive demand.

Discussion

With regard to factorial survey experiments, it is often argued that the presentation of information on the situational context when using vignettes allows to mirror the reality of everyday life because it leads to a more standardised imagination of the situation across respondents and less use of general heuristic principles by respondents, which will result in more reliable and valid responses (cf. Alexander & Becker, 1978; Armacost et al., 1991; Finch, 1987; Shamon et al., 2019). Previous studies emphasised the relevance of the presentation format of vignettes for response behaviour (e.g., Eifler & Petzold, 2014; Sauer et al., 2020; Shamon et al.,

2019). So far, there are hardly any empirical studies that have examined the implications of whether vignettes are presented in a written form or by means of visual stimuli, for example, by videos, photos or pictures. Therefore, we were concerned about possible differences between using written vignettes or photo vignettes in factorial survey experiments.

To pursue this question, we used a split ballot experiment employing two factorial survey experiments including either written vignettes or photo vignettes among the population of students from a German university. For our example of use, we referred to the broken windows theory, according to which signs of disorder lead to different levels of fear of crime in a scenario. We used vignettes that describe or display situations with varying signs of physical and social disorder serving as cues for the crime level in a situation. We randomly assigned respondents to one of the two presentation formats and asked them to report their perceived level of safety towards each vignette.

Following Shamon et al. (2019), we assumed that the response process for vignettes is characterised by a more unified retrieval of information from the memory. In line with this, we argued that, on the one hand, the standardised presentation of vignettes will evoke comparable interpretation frames in respondents but, on the other hand, differences between presenting either written vignettes or photo vignettes will occur. Referring to the DCT (Paivio, 1979; Sadoski & Paivio, 2013), we assumed that recall of situational information may depend on the presentation format used in factorial survey experiments. While the verbal information provided by written vignettes is processed in a sequential order, meaning that the verbal information has to be decoded first before visual mental representations are activated, information provided by photo vignettes is processed simultaneously, meaning that both verbal and visual representations are activated concurrently. In our example of use, the respondents were asked to report their feelings of unsafety with regard to everyday situations which provide varying cues of social disorder. We expected that the effects of physical and social signs of disorder should be stronger when presenting photo vignettes.

Our first result is that both vignette formats evoke almost identical distributions of self-reported feelings of unsafety and similar directions of the effects of all vignette dimensions. In accordance with the broken windows theory, cues for signs that indicate a high level of physical and/or social disorder mostly similarly increased the level of fear of crime expressed by the stated feelings of unsafety across both presentation formats. Especially the signs of social disorder, such as teenagers hanging around, show strong effects on the feelings of unsafety. In contrast, the location settings presented in the scenario show the least influence. This result indicates that both presentation formats seem similarly suited to evoke the retrieval of relevant cognitive and affective information from memory and to activate adequate mental representations for the interpretation of a situation.

Our second result is that the effects, conditional on the presentation format, differ significantly for the cues of physical disorder but not for the cues of social disorder. Respondents reacted more strongly to the lighting in a scenario and less to physical decay and littering in the photo vignettes. A reason for this result may be seen in the principal congruence between presenting visual information that particularly triggers the sense of vision as opposed to presenting visual information on physical or social aspects of a situation (see also Vanderveen, 2018). Whether this result may also indicate a higher level of validity of photo vignettes, however, cannot be answered on the basis of our study.

Additional residual analyses revealed that models of the written vignettes fit better when comparing between subjects, while there is a lack of differences when comparing answers of subjects within the presentation format. The mixed results do not indicate clear advantages of one presentation format over the other. Instead, one may gain some and also lose some by choosing either format.

Yet, the results do not fully correspond to our hypothesis, according to which all effects of the signs of disorder should be stronger in the photo vignettes. A possible explanation for the largely similar effect sizes across both presentation formats in the case of signs of social disorder is that they may have a strong activating effect also in the written vignettes. In contrast to signs of physical disorder, signs of social disorder may work as cues for potential social interaction. Humans are social beings and, therefore, may focus strongly on interacting with other people as part of their *conditio humana*. Through life-long learning of evaluating the social environment, people could develop a strong and inter-individually coherent representation of social situations. That is, when having conversations, or reading books or newspapers, people are trained to imagine other people and groups of people and their activities. This routine could lead to the consequence that social cues in both written vignettes and photo vignettes may evoke comparable cognitive processes so that the same effects are observed. This interpretation would be in line with the assumption that the response process for vignettes is characterised by a more unified retrieval of information from the memory (Shamon et al., 2019).

Moreover, in contradiction with our assumptions, a stronger effect of physical decay and littering on the reported feelings of unsafety was detected with the written vignettes. Yet, the assumption of a more unified retrieval of information from the memory allows for a coherent re-interpretation, as the stronger effect may reflect a methodological artefact. Compared to physical disorder in photo vignettes and to other dimensions in photo vignettes, more information is provided in written vignettes in this dimension. It covers two sentences including comprehensive information about a number of details (graffiti, beer bottles, garbage, wall plaster, asphalt holes). This may have overly activated the retrieval of related mental representations and evoked particular emotional arousal in the written vignettes, and

must be considered as a weakness of our design. Such problems should be avoided in future studies.

The results also contain implications for specific areas of the application of vignettes. First, though both modes seem to obtain meaningful effect estimates, applications with relevant visual information might eventually be better operationalised with photo vignettes. The reason may be that information which shall trigger the sense of vision may not be processed adequately when using written vignettes describing real-world scenarios. This is particularly the case when everyday situations are presented in which aspects of the physical environment are considered. Second, the results of this study also indicate that information on the social environment might be adequately processed also with written vignettes. Furthermore, there are many areas of application of vignettes in which the effects of the physical and, therefore, visible characteristics of a situation or of the physical characteristics of people are not theoretically significant. If no 'visual' dimensions are considered, photo vignettes will most likely not offer any advantage over written vignettes. Third, details may possibly be studied with the written format more comprehensively by actively drawing attention to specific dimensions through dense descriptions. This may appear as an advantage for special questions but may also be a disadvantage if certain factors are focused too strongly and over-activate certain attitudes or norms. In such a case, the importance of this dimension may be overestimated in written vignettes (see number of levels effect). However, photo vignettes may also be used to put a special emphasis on certain dimensions.

The main limitations of our study lie in the external validity of the results. The study was carried out with a homogeneous student sample from a single German university. Although this is sufficient for a method study of this kind, a replication with a more heterogeneous sample is desirable for the future, in order to ensure the robustness of the effects via subgroup analyses or cultural comparisons. In addition, the interpretation is limited by the varied vignette dimensions. It would be conceivable to use further or different operationalisations for social and physical cues in replications. It would also be promising to vary only social or only physical characteristics of the environment in order to increase the external validity. We also measured the stated feelings of unsafety as outcome variable, though the broken windows theory deals with the term fear of crime, which is not identical. Yet, we decided to measure the level of fear of crime using the standard indicator of this construct. The emotional dimension is brought to the fore with this outcome measurement. It would yet be interesting to measure not only attitudes towards the situation but also behavioural intentions, such as leaving the scenario or seeking protection.

As a conclusion, using visual stimuli might have advantages over written stimuli in research situations where visual information is theoretically relevant and particularly triggers the sense of vision. The reason for this is that visual stim-

uli allow to easily display spatial information concerning real-life situations that would be difficult to describe. Nevertheless, photo vignettes are reduced to a two-dimensional representation of aspects of social reality. This means that although they might work better in providing approximations to everyday life than written vignettes, they are not suitable for presenting additional sensory information like noise or smell (Mitchell, 1986). Furthermore, neither written vignettes nor photo vignettes are suitable for presenting information concerning bodily movement or other senses like, for example, the senses of touch, smell or hearing. While using video vignettes might solve some of these problems, as video clips may present both bodily movement and sound, other aspects of real-life situations might possibly only be simulated by means of methods involving virtual realities (Van Gelder et al., 2019; Van Sintemaartensdijk et al., 2020). Future studies concerning the relevance of the presentation format of vignettes should take these considerations into account.

References

- Alexander, C. S., & Becker, H. J. (1978). The Use of Vignettes in Survey Research. *Public Opinion Quarterly*, 42(1), 93-104. <https://doi.org/10.1086/268432>
- Alexander, N. (2015). What's More General than a Whole Population? *Emerging Themes in Epidemiology*, 12, 11-11. <https://doi.org/10.1186/s12982-015-0029-4>
- Armocost, R. L., Hosseini, J. C., Morris, S. A., & Rehbein, K. A. (1991). An Empirical Comparison of Direct Questioning, Scenario, and Randomized Response Methods for Obtaining Sensitive Business Information. *Decision Sciences*, 22(5), 1073-1090. <https://doi.org/10.1111/j.1540-5915.1991.tb01907.x>
- Auspurg, K., & Hinz, T. (2015). *Factorial Survey Experiments*. Thousand Oaks: Sage.
- Barthes, R. (1977). *Image, Music, Text*. London: Fontana Press.
- Beck, M., & Opp, K.-D. (2001). Der faktorielle Survey und die Messung von Normen. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 53(2), 283-306. <https://doi.org/10.1007/s11577-001-0040-3>
- Benson, L. E. (1941). Studies in Secret-Ballot Technique. *Public Opinion Quarterly*, 5, 79-82. <https://doi.org/10.1086/265464>
- Berk, R. A., Western, B., & Weiss, R. E. (1995). Statistical Inference for Apparent Populations. *Sociological Methodology*, 25, 421-458. <https://doi.org/10.2307/271073>
- Bower, G. H. (1970). Imagery as a Relational Organizer in Associative Learning. *Journal of Verbal Learning and Verbal Behavior*, 9, 529-533. [https://doi.org/10.1016/s0022-5371\(70\)80096-2](https://doi.org/10.1016/s0022-5371(70)80096-2)
- Brockmole, J. R. (2009). *The Visual World in Memory*. Hove, Sussex: Psychology Press.
- Caro, F. G., Ho, T. H., McFadden, D., Gottlieb, A. S., Yee, C., Chan, T., & Winter, J. (2012a). Using the Internet to Administer More Realistic Vignette Experiments. *Social Science Computer Review*, 30(2), 184-201. <https://doi.org/10.1177/0894439310391376>

- Caro, F. G., Yee, C., Levien, S., Gottlieb, A. S., Winter, J., McFadden, D. L., & Ho, T. H. (2012b). Choosing Among Residential Options: Results of a Vignette Experiment. *Research on Aging, 34*(1), 3-33. <https://doi.org/10.1177/0164027511404032>
- Collier Jr, J. (1957). Photography in Anthropology: A Report on Two Experiments. *American Anthropologist, 59*(5), 843-859. <https://doi.org/10.1525/aa.1957.59.5.02a00100>
- Collier, J., & Collier, M. (1986). *Visual Anthropology: Photography as a Research Method*. Albuquerque: University of New Mexico Press.
- Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009). *Internet, Mail, and Mixed-Mode Surveys. The Tailored Design Method*. Wiley: New Jersey.
- Dinora, P., Schoeneman, A., Dellinger-Wray, M., Cramer, E. P., Brandt, J., & D'Aguilar, A. (2020). Using Video Vignettes in Research and Program Evaluation for People with Intellectual and Developmental Disabilities: A Case Study of the Leadership for Empowerment and Abuse Prevention (LEAP) Project. *Evaluation and Program Planning, 79*. <https://doi.org/10.1016/j.evalprogplan.2019.101774>
- Edgington, E. (1966). Statistical Inference and Nonrandom Samples. *Psychological Bulletin, 66*(6), 485-487. <https://doi.org/10.1037/h0023916>
- Eifler, S. (2007). Evaluating the Validity of Self-Reported Deviant Behavior Using Vignette Analyses. *Quality & Quantity, 41*, 303-318. <https://doi.org/10.1007/s11135-007-9093-3>
- Eifler, S., & Petzold, K. (2014). Der Einfluss der Ausführlichkeit von Vignetten auf die Erfassung prosozialer Einstellungen. Ergebnisse zweier Split-Ballot Experimente. *Soziale Welt. Zeitschrift für sozialwissenschaftliche Forschung und Praxis, 2014*(2), 247-270. <https://doi.org/10.5771/0038-6073-2014-2-247>
- Eifler, S., Petzold, K., & Verbeek-Teres, M. (2021). Presentation Format Differences in Factorial Surveys. *Eichstätter Beiträge zur Soziologie, 19*. Retrieved December 13, 2021, from <https://edoc.ku.de/id/eprint/29215/>
- Finch, J. (1987). Research Note. The Vignette Technique in Survey Research. *Sociology, 21*(1), 105-114. <https://doi.org/10.1177/0038038587021001008>
- Findlay, J. M., & Gilchrist, I. D. (2003). *Active Vision: The Psychology of Looking and Seeing*. Oxford: Oxford University Press.
- Golden III, J. H., Johnson, C. A., & Lopez, R. A. (2001). Sexual Harassment in the Workplace: Exploring the Effects of Attractiveness on Perception of Harassment. *Sex Roles, 45*(11/12), 767-784.
- Goyal, N., Wice, M., Kinsbourne, M., & Castano, E. (2017). A Picture Is Worth a Thousand Words: The Influence of Visuospatial and Verbal Cognitive Styles on Empathy and Willingness to Help. *Social Psychology, 48*(6) 372-379. <https://doi.org/10.1027/1864-9335/a000318>
- Hanna, A., & Loftus, G. (1993). A Model for Conceptual Processing of Naturalistic Scenes. *Canadian Journal of Experimental Psychology, 47*(3), 548-569. <https://doi.org/10.1037/h0078851>
- Harper, D. (2002). Talking about Pictures: A Case for Photo Elicitation. *Visual Studies, 17*(1), 13-26. <https://doi.org/10.1080/14725860220137345>
- Havekes, E., Coenders, M., & Van der Lippe, T. (2013). Positive or Negative Ethnic Encounters in Urban Neighbourhoods? A Photo Experiment on the Net Impact of Ethnicity and Neighbourhood Context on Attitudes Towards Minority and Majority Residents. *Social Science Research, 42*(4), 1077-1091. <https://doi.org/10.1016/j.ssresearch.2013.02.002>

- Hox, J. J., Kreft, I. G. G., & Hermkens, P. L. J. (1991). The Analysis of Factorial Surveys. *Sociological Methods & Research, 19*(4), 493-510. <https://doi.org/10.1177/0049124191019004003>
- Hughes, R., & Huby, M. (2004). The Construction and Interpretation of Vignettes in Social Research. *Social Work and Social Sciences Review, 11*(1), 36-51. <https://doi.org/10.1921/17466105.11.1.36>
- Hunter, A. (1978). *Symbols of Incivility*. Paper presented at the Annual Meeting of the American Society of Criminology, November 1978, Dallas.
- Jasso, G. (2006). Factorial Survey Methods for Studying Belief and Judgments. *Sociological Methods & Research, 34*(3), 334-423. <https://doi.org/10.1177/0049124105283121>
- Jenkins, J. R., Neale, D. C., & Deno, S. L. (1967). Differential Memory for Picture and Word Stimuli. *Journal of Educational Psychology, 58*(5), 303-307. <https://doi.org/10.1037/h0025025>
- Keizer, K., Lindenberg, S., & Steg, L. (2014). Doing Field Studies: What is it All About? *Group Processes & Intergroup Relations, 17*(3), 404-410. <https://doi.org/10.1177/1368430213510750>
- Kelling, G. L., & Coles, C. M. (1996). *Fixing Broken Windows: Restoring Order and Reducing Crime in Our Communities*. New York: Touchstone.
- Keuschnigg, M., & Wolbring, T. (2015). Disorder, Social Capital, and Norm Violation. Three Field Experiments on the Broken Windows Thesis. *Rationality and Society, 27*(1), 96-126. <https://doi.org/10.1177/1043463114561749>
- Kosslyn, S. M. (1981). The Medium and the Message in Mental Imagery: A Theory. *Psychological Review, 88*(1), 46-66. <https://doi.org/10.1037/0033-295x.88.1.46>
- Kosslyn, S. M., & Pomerantz, J. R. (1977). Imagery, Propositions, and the Form of Internal Representations. *Cognitive Psychology, 9*, 52-76. [https://doi.org/10.1016/0010-0285\(77\)90004-4](https://doi.org/10.1016/0010-0285(77)90004-4)
- Krysan, M., Couper, M. P., Farley, R., & Forman, T. A. (2009). Does Race Matter in Neighborhood Preferences? Results from a Video Experiment. *American Journal of Sociology, 115*(2), 527-559. <https://doi.org/10.1086/599248>
- Lewis, D. A., & Salem, G. (1986). *Fear of Crime. Incivility and the Production of a Social Problem*. New Brunswick, NJ: Transaction Books.
- Liebig S., Sauer, C., & Friedhoff, S. (2015). Empirische Gerechtigkeitsforschung mit dem faktoriellen Survey. In M. Keuschnigg, & T. Wolbring (Eds.), *Experimente in den Sozialwissenschaften* (pp 321-339). Baden-Baden: Nomos.
- Luck, S. J., & Hollingworth, A. (2008). *Visual Memory*. Oxford: Oxford University Press.
- Manghani, S. (2013). *Image Studies. Theory and Practice*. London, New York: Routledge.
- Marion, J. S., & Crowder, J. W. (2013). *Visual Research. A Concise Introduction to Thinking Visually*. London: Bloomsbury.
- McCloud, S. (1994). *Understanding Comics. The Invisible Art*. New York: Harper Perennial.
- Mitchell, W. J. T. (1986). *Iconology. Image, Text, Ideology*. Chicago, London: The University of Chicago Press.
- Mutz, D. C. (2011). *Population-Based Survey Experiments*. Princeton, NJ: Princeton University Press.

- O'Donnell, A. B., Lutfey, K. E., Marceau, L. D., & McKinlay, J. B. (2007). Using Focus Groups to Improve the Validity of Cross-National Survey Research: A Study of Physician Decision Making. *Qualitative Health Research, 17*(7), 971–981. <https://doi.org/10.1177/1049732307305257>
- Paivio, A. (1979). *Imagery and Verbal Processes*. New Jersey: Lawrence Erlbaum Associates.
- Piquero, A. (1999). The Validity of Incivility Measures in Public Housing. *Justice Quarterly, 16*(4), 793–818. <https://doi.org/10.1080/07418829900094371>
- Rashotte, L. S. (2003). Written Versus Visual Stimuli in the Study of Impression Formation. *Social Science Research, 32*(2), 278-293. [https://doi.org/10.1016/s0049-089x\(02\)00050-9](https://doi.org/10.1016/s0049-089x(02)00050-9)
- Rose, G. (2012). *Visual Methodologies. An Introduction to Researching with Visual Materials*. London: Sage Publications.
- Rossi, P. H. (1979). Vignette Analysis: Uncovering the Normative Structure of Complex Judgements. In R. K. Merton, J. S. Coleman, & P. H. Rossi (Eds.), *Qualitative and Quantitative Social Research: Papers in Honor of Paul F. Lazarsfeld* (pp. 176–186). New York, NY: Free Press.
- Rossi, P. H., & Anderson, A. B. (1982). The Factorial Survey Approach: An Introduction. In P. H. Rossi, & S. L. Nock (Eds.), *Measuring Social Judgments: The Factorial Survey Approach* (pp 15-67). Beverly Hills, CA: Sage Publications.
- Sadoski, M., & Paivio, A. (2013). *Imagery and Text: A Dual Coding Theory of Reading and Writing*. London: Routledge.
- Sauer, C., Auspurg, K., & Hinz, T. (2020). Designing Multi-Factorial Survey Experiments: Effects of Presentation Style (Text or Table), Answering Scales, and Vignette Order. *methods, data, analyses, 14*(2), 195-214. <https://doi.org/10.12758/mda.2020.06>
- Schwarzbach, H., & Eifler, S. (2020). Einflüsse der Präsentationsform eines faktoriellen Surveys zur Erfassung von Sicherheit im öffentlichen Raum. *Eichstätter Beiträge zur Soziologie, 18*. Retrieved December 13, 2021, from <https://edoc.ku.de/id/eprint/25734/>
- Shamon, H., Dülmer, H., & Giza, A. (2019). The Factorial Survey: The Impact of the Presentation Format of Vignettes on Answer Behavior and Processing Time. *Sociological Methods & Research* (Onlinefirst). <https://doi.org/10.1177/0049124119852382>
- Shepard, R. N. (1967). Recognition Memory for Words, Sentences, and Pictures. *Journal of Verbal Learning and Verbal Behavior, 6*, 156-163. [https://doi.org/10.1016/s0022-5371\(67\)80067-7](https://doi.org/10.1016/s0022-5371(67)80067-7)
- Skogan, W. G. (1978). *Victimization Surveys and Criminal Justice Planning*. Washington, DC: National Institute of Law Enforcement and Criminal Justice.
- Skogan, W. G. (1990). *Disorder and Decline. Crime and the Spiral Decay in American Neighborhoods*. New York, NY: Free Press.
- Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. London: Sage.
- Taylor, R. B. (1999). The Incivilities Thesis: Theory, Measurement, and Policy. In R. H. Langworthy (Eds.), *Measuring what Matters: Proceedings from the Policing Research Institute Meetings* (pp 65–88). Washington, DC: U.S. Department of Justice.
- Tourangeau, R. (1984). Cognitive Science and Survey Methods. In T. B. Jabine, M. L. Straf, J. M. Tanur, & R. Tourangeau (Eds.), *Cognitive Aspects of Survey Design: Building a Bridge Between Disciplines* (pp 73–100). Washington, DC: National Academy Press.

-
- Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge: Cambridge University Press.
- Van Gelder, J.-L., De Vries, R. E., Demetriou, A., Van Sintemaartensdijk, I., & Donker, T. (2019). The Virtual Reality Scenario Method: Moving from Imagination to Immersion in Criminal Decision-Making Research. *Journal of Research in Crime and Delinquency*, 56(3), 451–480. <https://doi.org/10.1177/0022427818819696>
- Van Sintemaartensdijk, I., Van Gelder, J.-L., Van Prooijen, J.-W., Nee, C., Otte, M., & Van Lange, P. (2020). Mere Presence of Informal Guardians Deters Burglars: A Virtual Reality Study. *Journal of Experimental Criminology*. <https://doi.org/10.1007/s11292-020-09430-1>
- Vanderveen, G. (2018). Visual Methods in Research on Fear of Crime: A Critical Assessment. In M. Lee, & G. Mythen (Eds.), *The Routledge International Handbook on Fear of Crime* (pp 170-189). London: Routledge. <https://doi.org/10.4324/9781315651781-13>
- Wallander, L. (2009). 25 Years of Factorial Surveys in Sociology: A Review. *Social Science Research*, 38, 505-520. <https://doi.org/10.1016%2Fj.ssresearch.2009.03.004>
- Wilson, J. Q., & Kelling, G. L. (1982). Broken Windows: The Police and Neighborhood Safety. *Atlantic Monthly*, 249(3), 29-38.

Appendix

Table A-1 Robustness regarding evaluation of realistic vignette descriptions

Reported feelings of unsafety	Evaluation: realistic and non-realistic		Evaluation: only realistic	
Pedestrian underpass (ref. wide square)	0.173***	(14.06)	0.185***	(13.85)
Teenagers hanging around (ref. couple goes for a walk)	0.755***	(61.44)	0.775***	(58.12)
Physical decay / littering (ref. no)	0.295***	(20.50)	0.306***	(19.60)
No passers-by (ref. adult passers-by)	0.295***	(22.81)	0.314***	(22.31)
Gloomy, video surveillance (ref. lighting)	0.382***	(26.18)	0.387***	(24.53)
Gloomy, no surveillance (ref. lighting)	0.427***	(27.68)	0.432***	(25.78)
Photo vignette (ref. written vignette)	0.0155	(0.56)	0.00672	(0.22)
Vignette set (ref. set 1)				
Set 2	-0.0629	(-1.33)	-0.0672	(-1.29)
Set 3	-0.190***	(-3.89)	-0.191***	(-3.55)
Set 4	-0.214***	(-4.40)	-0.222***	(-4.14)
Set 5	-0.200***	(-4.06)	-0.214***	(-4.00)
Set 6	-0.0896	(-1.81)	-0.0593	(-1.08)
Constant	0.259***	(6.21)	0.228***	(4.95)
σ_u	0.432		0.436	
σ_c	0.571		0.569	
Log likelihood	-9,293.0		-7,838.9	
LR- χ^2	4,453.45***		3,948.75***	
AIC	18,615.95		15,707.84	
$N_{\text{Vignettes}}$	9,620		8,129	
$N_{\text{Respondents}}$	1,207		1,019	

Linear random intercept maximum likelihood estimations.

t statistics in parentheses; * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$

Table A-2 Robustness with and without covariates

Reported feelings of unsafety	Without covariates		With covariates	
Pedestrian underpass (ref. wide square)	0.184***	(13.69)	0.184***	(13.55)
Teenagers hanging around (ref. couple goes for a walk)	0.776***	(57.82)	0.780***	(57.33)
Physical decay / littering (ref. no)	0.310***	(20.47)	0.311***	(19.52)
No passers-by (ref. adult passers-by)	0.305***	(21.81)	0.319***	(22.26)
Gloomy, video surveillance (ref. lighting)	0.375***	(23.47)	0.382***	(23.77)
Gloomy, no surveillance (ref. lighting)	0.432***	(25.70)	0.434***	(25.38)
Photo vignette (ref. written vignette)	0.0165	(0.53)	0.00768	(0.25)
Vignette set (ref. set 1)				
Set 2			-0.0902	(-1.77)
Set 3			-0.212***	(-3.99)
Set 4			-0.240***	(-4.56)
Set 5			-0.227***	(-4.33)
Set 6			-0.0533	(-0.99)
Age			-0.00206	(-0.57)
Male (ref. female)			-0.289***	(-9.09)
Spouse (ref. no partnership)			-0.0121	(-0.39)
Evaluation: questionnaire fatiguing			0.00143	(0.04)
Constant	0.107***	(3.73)	0.400***	(3.63)
σ_u	0.446		0.416	
σ_e	0.570		0.570	
Log likelihood	-7,605.4		-7,547.3	
LR- χ^2	3,806.7***		3,922.76***	
AIC	15,230.7		15,132.7	
$N_{\text{Vignettes}}$	7,855		7,855	
$N_{\text{Respondents}}$	985		985	

Linear random intercept maximum likelihood estimations.

t statistics in parentheses; * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$

Table A-3 Robustness by linear / ordered logit multi-level modelling

Reported feelings of unsafety	Linear multi-level model		Ord. log. multi-level model	
Pedestrian underpass (ref. wide square)	0.185***	(13.85)	0.657***	(13.07)
Teenagers hanging around (ref. couple goes for a walk)	0.775***	(58.12)	2.877***	(46.48)
Physical decay / littering (ref. no)	0.306***	(19.60)	1.154***	(19.29)
No passers-by (ref. adult passers-by)	0.314***	(22.31)	1.171***	(21.66)
Gloomy, video surveillance (ref. lighting)	0.387***	(24.53)	1.454***	(23.52)
Gloomy, no surveillance (ref. lighting)	0.432***	(25.78)	1.618***	(24.76)
Photo vignette (ref. written vignette)	0.00672	(0.22)	0.0266	(0.23)
Vignette set (ref. set 1)				
Set 2	-0.0672	(-1.29)	-0.301	(-1.51)
Set 3	-0.191***	(-3.55)	-0.650**	(-3.17)
Set 4	-0.222***	(-4.14)	-0.827***	(-4.03)
Set 5	-0.214***	(-4.00)	-0.799***	(-3.90)
Set 6	-0.0593	(-1.08)	-0.221	(-1.05)
Constant	0.228***	(4.95)		
Cut 1			1.249***	(7.07)
Cut 2			4.928***	(26.20)
Cut 3			7.897***	(38.48)
σ_u			2.812***	(16.24)
σ_u	0.436		2.812	
σ_c	0.569			
Log likelihood	-7,838.9		-7,495.9	
LR/Wald- χ^2	3,948.75***		2,766.59***	
AIC	15,707.84		15,023.84	
$N_{\text{Vignettes}}$	8,129		8,129	
$N_{\text{Respondents}}$	1,019		1,019	

Linear random intercept maximum likelihood estimation & ordered logit random intercept maximum likelihood estimation.

t statistics in parentheses; * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$

Table A-4 Robustness by linear / logit multi-level modelling

Reported feelings of unsafety	Linear multi-level model		Log. multi-level model	
Pedestrian underpass (ref. wide square)	0.185***	(13.85)	0.0906***	(11.24)
Teenagers hanging around (ref. couple goes for a walk)	0.775***	(58.12)	0.356***	(50.06)
Physical decay / littering (ref. no)	0.306***	(19.60)	0.169***	(18.36)
No passers-by (ref. adult passers-by)	0.314***	(22.31)	0.154***	(19.29)
Gloomy, video surveillance (ref. lighting)	0.387***	(24.53)	0.154***	(16.53)
Gloomy, no surveillance (ref. lighting)	0.432***	(25.78)	0.183***	(18.38)
Photo vignette (ref. written vignette)	0.00672	(0.22)	0.00745	(0.53)
Vignette set (ref. set 1)				
Set 2	-0.0672	(-1.29)	0.0101	(0.41)
Set 3	-0.191***	(-3.55)	-0.0394	(-1.62)
Set 4	-0.222***	(-4.14)	-0.0467	(-1.89)
Set 5	-0.214***	(-4.00)	-0.0605*	(-2.54)
Set 6	-0.0593	(-1.08)	0.0276	(1.02)
Constant	0.228***	(4.95)		
σ_u	0.436		1.648	
σ_e	0.569			
Log likelihood	-7,838.9		-3,513.5	
LR/Wald- χ^2	3,948.75***		1,392.5***	
AIC	15,707.84		7,055.06	
$N_{\text{Vignettes}}$	8,129		8,129	
$N_{\text{Respondents}}$	1,019		1,019	

Linear random intercept maximum likelihood estimation & logistic random intercept maximum likelihood estimation.

t statistics in parentheses; * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$

Note: Logistic multi-level model with median split of dependent variable; Average Marginal Effects reported (AME).

Table A-5 Robustness by multi-level structure

Reported feelings of unsafety	Linear multi-level model		Linear OLS model	
Pedestrian underpass (ref. wide square)	0.185***	(13.85)	0.185***	(13.02)
Teenagers hanging around (ref. couple goes for a walk)	0.775***	(58.12)	0.776***	(43.22)
Physical decay / littering (ref. no)	0.306***	(19.60)	0.306***	(17.48)
No passers-by (ref. adult passers-by)	0.314***	(22.31)	0.315***	(20.83)
Gloomy, video surveillance (ref. lighting)	0.387***	(24.53)	0.387***	(20.78)
Gloomy, no surveillance (ref. lighting)	0.432***	(25.78)	0.432***	(22.08)
Photo vignette (ref. written vignette)	0.00672	(0.22)	0.0115	(0.38)
Vignette set (ref. set 1)				
Set 2	-0.0672	(-1.29)	-0.0675	(-1.29)
Set 3	-0.191***	(-3.55)	-0.191***	(-3.43)
Set 4	-0.222***	(-4.14)	-0.222***	(-4.13)
Set 5	-0.214***	(-4.00)	-0.215***	(-3.80)
Set 6	-0.0593	(-1.08)	-0.0585	(-1.07)
Constant	0.228***	(4.95)	0.224***	(4.73)
σ_u	0.436			
σ_c	0.569			
Log likelihood	-7,838.9		-8,823.8	
LR- χ^2 / F	3,948.75***		318.10***	
AIC	15,707.84		17,673.65	
$N_{\text{Vignettes}}$	8,129		8,129	
$N_{\text{Respondents}}$	1,019			

Linear random intercept maximum likelihood estimation & linear ordinary least squares estimations, clustered SE.

t statistics in parentheses; * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$

Table A-6 Estimation of interactions between presentation format and vignette dimensions

Reported feelings of unsafety		
Photo vignette (ref. written vignette)	0.0958	(1.13)
Pedestrian underpass (ref. wide square)	0.200***	(10.35)
Teenagers hanging around (ref. couple goes for a walk)	0.778**	(40.41)
Physical decay / littering (ref. no)	0.395**	(17.48)
No passers-by (ref. adult passers-by)	0.333**	(16.38)
Gloomy, video surveillance (ref. lighting)	0.217**	(9.47)
Gloomy, no surveillance (ref. lighting)	0.268**	(11.05)
Vignette set (ref. set 1)		
Set 2	0.00311	(0.04)
Set 3	-0.00650	(-0.08)
Set 4	-0.174*	(-2.23)
Set 5	-0.0264	(-0.34)
Set 6	0.0332	(0.42)
Photo vignette * Underpass	-0.0288	(-1.09)
Photo vignette * Teenagers hanging around	-0.00468	(-0.18)
Photo vignette * Physical decay / littering	-0.167**	(-5.40)
Photo vignette * No passers-by	-0.0402	(-1.44)
Photo vignette * Gloomy, video surveillance	0.318**	(10.16)
Photo vignette * Gloomy, no surveillance	0.309**	(9.30)
Photo vignette * set 2	-0.131	(-1.27)
Photo vignette * set 3	-0.346**	(-3.24)
Photo vignette * set 4	-0.0834	(-0.79)
Photo vignette * set 5	-0.349**	(-3.28)
Photo vignette * set 6	-0.165	(-1.53)
Constant	0.181**	(2.87)
σ_u	0.431	
σ_c	0.564	
Log likelihood	-7,755.1876	
LR- χ^2	4,116.22***	
$N_{\text{Vignettes}}$	8,129	
$N_{\text{Respondents}}$	1,019	

Linear random intercept maximum likelihood estimations.

t statistics in parentheses; * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$

Note: Conditional effects for presentation format of written vignettes.

Table A-7 Estimation of dimensions' effects by presentation format

Reported feelings of unsafety	Written vignettes		Photo vignettes	
Pedestrian underpass (ref. wide square)	0.200***	(10.68)	0.171***	(9.18)
Teenagers hanging around (ref. couple goes for a walk)	0.778***	(41.68)	0.773***	(41.50)
Physical decay / littering (ref. no)	0.395***	(18.04)	0.227***	(10.45)
No passers-by (ref. adult passers-by)	0.333***	(16.90)	0.293***	(14.93)
Gloomy, video surveillance (ref. lighting)	0.217***	(9.77)	0.536***	(24.41)
Gloomy, no surveillance (ref. lighting)	0.268***	(11.40)	0.577***	(24.64)
Photo vignette (ref. written vignette)	0.00672	(0.22)	0.0115	(0.38)
Vignette set (ref. set 1)				
Set 2	0.00315	(0.04)	-0.128	(-1.84)
Set 3	-0.00647	(-0.08)	-0.353***	(-4.84)
Set 4	-0.174*	(-2.24)	-0.257***	(-3.54)
Set 5	-0.0263	(-0.34)	-0.375***	(-5.19)
Set 6	0.0332	(0.43)	-0.132	(-1.76)
Constant	0.181**	(2.90)	0.277***	(4.79)
σ_u	0.431***		0.430***	
σ_c	0.546***		0.578***	
Log likelihood	-3,534.8		-4,214.6	
LR- χ^2	1,894.0***		2,212.3***	
AIC	7,097.508		8,457.291	
$R^2_{\text{within}}^+$	0.414		0.435	
$R^2_{\text{between}}^+$	0.206		0.102	
$R^2_{\text{overall}}^+$	0.332		0.326	
$N_{\text{Vignettes}}$	3,816		4,313	
$N_{\text{Respondents}}$	479		540	

Linear random intercept maximum likelihood estimations.

+ from linear random intercept generally least squares estimations.

t statistics in parentheses; * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$

The Impact of Presentation Format on Conjoint Designs: A Replication and an Extension

Sophie Cassel, Josefine Magnusson & Sebastian Lundmark
University of Gothenburg

Abstract

In recent years, conjoint experiments have been in vogue across the social sciences. A reason for the conjoint experiments' popularity is that they allow researchers to estimate the causal effects of many components of stimuli simultaneously. However, for conjoint experiments to produce valid results, respondents need to be able to process and understand the wide range of dimensions presented to them in the experiment. If the information processing is too demanding or too complicated, respondents are likely to turn to satisficing strategies, leading to poorer data quality and subsequently decreasing the researcher's ability to make accurate causal inferences. One factor that may lead to the adoption of satisficing strategies is the presentation format used for the conjoint experiment (i.e., presenting the information within a text paragraph or a table). In the present paper, a direct replication of the single conjoint presentation format experiment described in Shamon, Dülmer, and Giza's (2019) paper in *Sociological Methods & Research* is presented, and extending their work to paired conjoint experiment. The results of the direct replication showed that respondents evaluated the questionnaire more favorably when reading the table format but were, on the other hand, less likely to participate in subsequent panel waves. Albeit the number of break-offs, refusals, and non-responses did not differ between the two formats, respondents who saw the table format evaluated the scenarios with more consistency and less dimension reduction, thus favoring the table presentation format. For paired conjoint experiments, the presentation format did not affect survey evaluations or panel participation but the table format heavily outperformed the text format on every data quality measure except for dimension reduction. Conceptually, albeit not directly replicating the findings in Shamon, Dülmer, and Giza (2019), the present manuscript concludes that the table format appears preferable over the text format for conjoint experimental designs.

Keywords: conjoint experiments, satisficing, presentation format, text versus tables, replication



Conjoint experiments (also referred to as “factorial surveys”) are widely used in the social sciences for measuring beliefs and preferences, and for multidimensional decision making (Shamon, Dülmer & Giza, 2019). In a typical conjoint experiment, respondents are asked to evaluate a vignette with several different dimensions (i.e., attributes of the sets/profiles to evaluate) with randomly assigned levels (i.e., the value of the dimension, for example, male/female, rich/poor).

An advantage of designs such as conjoint experiments is that the evaluation processes of the vignette resemble the judgment processes made in real life given that it involves making a single evaluation based on the information for several attributes at the same time (Hainmueller, Hopkins & Yamamoto, 2014). In comparison to the single factor experiment, the controlled variation of attribute levels in the conjoint experiment enables researchers to capture the unique impact of several dimensions at the same time (Auspurg & Jäckle, 2017; Hainmueller et al., 2014). By providing respondents with the relevant information needed to form an opinion, the conjoint experiment is thought to reduce the potentiality for respondents to change their reference frame when answering a survey (Shamon et al., 2019). Consequently, these potential benefits, combined with the growing practice of administering questionnaires via computers, have made the conjoint experiment popular across many research fields.

Despite the increasing popularity of the conjoint experiment, the number of studies examining the effect of design complexity (e.g., investigating how the vignette is presented) on respondents’ cognitive burden and answer behavior has been relatively few (see Shamon et al., 2019). Except for Shamon, Dülmer, and Giza (2019), Sauer, Auspurg and Hinz (2020), and Hainmueller, Hangartner, and Yamamoto (2015), there are no studies on the impact that the presentation formats may have on the cognitive burden and answer behavior of respondents and the effect is still relatively unknown. As information intake is central to vignette and conjoint studies, knowledge of how respondents react to different presentation formats is essential for researchers’ ability to draw accurate inferences from the experiments (Shamon et al., 2019).

Two presentation formats dominate the realm of online conjoint experiments: text (presenting the information in a text paragraph) and table format (a table with dimensions and levels in rows and columns). Other formats exist, such as presenting the information via video clips, illustrated cards, or pictures (see, e.g., Sato, Kubo & Namatame, 2007), but the present study focuses on text and table formats. Theoretically, using a text format may be preferable because nesting the information in stories may enhance respondents’ understanding of the hypothetical situa-

Direct correspondence to

Sophie Cassel, the SOM Institute, University of Gothenburg, Sweden
E-mail: Sophie.Cassel@gu.se

tion or increase empathy with the described situation. The enhanced understanding and empathy may, in turn, increase the respondents' attention to the dimensions in the conjoint experiment (Auspurg & Hinz, 2015). Furthermore, individuals may be more accustomed to absorbing information in flowing text than from reading tables. Given respondents' likelihood of being more used to reading text than tables, presenting conjoint experiments in text format may make respondents less prone to *satisficing strategies*. Satisficing strategies are employed when respondents resort to suboptimal information processing and response techniques as a means to lower their cognitive burden of filling out a questionnaire or participating in an experiment (Krosnick, 1999).

By contrast, Shamon et al. (2019, p. 9) argued that table formats may facilitate stronger information intake in comparison to text formats because tables, generally, contain fewer words and present the relevant information at the same visual position across different scenarios, which should decrease the cognitive investment needed by respondents and especially among the respondents with lower cognitive skills (Shamon et al., 2019). However, Shamon et al.'s (2019) assumption only hold if respondents accurately comprehend the order (the rows and columns) in which the tables should be read. Such an assumption is more likely to be violated when respondents are asked to compare two different sets of characteristics (e.g., stating a preference for one of two political candidates with different dimensions), meaning that the respondents have to understand that the values in one column belong to the first set or profile (e.g., candidate A) and the values in the second column belong to the second (e.g., candidate B).

The existing empirical studies on the effect of presentation format on answering behavior paint an inconclusive picture. Whereas Sauer et al. (2020) found that table formats produced similar evaluations as text formats, Hainmueller et al. (2015) found that table format (evaluation of two sets/profiles) performed better than text formats in predicting real-life judgments. Similarly, Shamon et al. (2019) found support for the table format performing slightly better in reducing satisficing behavior compared to the text formats, although most of their measurements showed no statistically significant differences between the two formats. It appears that more data is needed to assess the role that the presentation format may have for the data quality and the respondents' experiences when participating in conjoint experiments.

To that end, the present paper presents a replication of Shamon et al.'s (2019) study on single conjoint experiments and the impact that the presentation format (text vs. table) may have on respondents' reporting behavior and subjective experience of the questionnaire is assessed. Following the advice of Sauer et al. (2020), the present study replicated Shamon et al.'s (2019) experiment in a non-student sample invited to resemble the Swedish population in terms of age, sex, and education. The sample was drawn from the Swedish Citizen Panel administered by the

Laboratory of Opinion Research (LORE) at the University of Gothenburg. Furthermore, extending Shamon et al.'s (2019) experiment, a conceptual replication of their study was performed, testing the impact of the presentation format in a paired conjoint setting (i.e., where respondents are asked to state their preferences regarding two different sets/profiles, for example by reporting their preference for one of two politicians described in the vignette). Furthermore, the importance of replicating published research has become especially acute given the many recent failures to replicate published literature (e.g., Open Science Collaboration, 2015).

This paper is organized as follows: First, a brief theoretical rationale for how text and table presentation formats may influence data quality is presented. Then, the hypotheses for the single conjoint and, thereafter, the paired conjoint experiment is introduced. The paper continues with a description of the evaluation criteria, methods and materials. The results are thereafter presented for the single conjoint and the paired conjoint experiment, separately. Next, a summary of the results and a comparison to Shamon et al.'s (2019) findings are presented. The paper ends with a discussion and some conclusions of the main takeaways of the paper.

How Does the Presentation Format Influence Data Quality and Respondent Behavior?

Single Conjoint Experiment

Survey respondents are sensitive to a range of, sometimes almost undetectable, survey design features (Schuman & Presser, 1996; Roberts et al., 2019). Previous research has suggested that the visual appearance of a questionnaire may influence respondents' satisfaction with it (although this connection is not always found – see, e.g., Mahon-Haft & Dillman, 2010). Furthermore, people have been found to interpret a task with a difficult-to-read instruction as more difficult to complete compared to when the task is described in an easy-to-read instruction (e.g., the information is presented in an easy- or difficult-to-read font) (Song & Schwarz, 2008). Hence, it stands to reason that the presentation format of a conjoint experiment that is more difficult to read, interpret, and time-consuming for the respondent, may produce less overall satisfaction with the questionnaire and may make respondents more likely to rate the questionnaire as difficult to complete.

In the presentation format of conjoint experiments, presenting the information in text format generally requires a greater number of characters and syllables, and there are more words to read compared to when the information is presented in table form. The fewer number of words and characters on the screen in a table may be interpreted as less information to process for the respondent and, therefore, as a less demanding task, potentially leading to greater respondent satisfaction com-

pared to when the information is presented as text. On the other hand, respondents may be more used to gathering information in text format than in table format, which may lead to greater survey satisfaction with the text format. Therefore, the following hypothesis will be assessed:

Hypothesis 1a (H1a): Respondents who evaluate scenarios presented in a table format may report a better respondent experience (i.e., report greater satisfaction with the questionnaire, take less time to evaluate the scenarios, and be more likely to participate in future panel waves) than respondents who evaluate the same scenarios presented as text.

According to satisficing theory, a presentation format that is difficult to understand, hard to process, or difficult to read is thought to induce stronger satisficing behavior (Krosnick & Alwin, 1987; Krosnick, 1991; Song & Schwartz, 2008). Satisficing is a decision-making process in which a person, instead of expending appropriate cognitive effort to come to an optimum decision, decides to expend only the minimum effort needed (or no effort at all) to come to a decision (Simon, 1957; Krosnick, 1991). Compared to expending the appropriate effort, satisficing strategies rarely lead to decisions or answers that best represent an individual's actual wants, needs, or attitudes. A presentation format in conjoint experiments that increases the likelihood of a respondent satisficing is expected to produce a range of negative influences on data quality.

When presenting conjoint experiments in a table format, the researcher's goal is to make the necessary information easily accessible to the respondent by only presenting the information needed for the respondent to make a decision, and to present that information through a minimum number of characters, syllables, and words, and using an easy-to-read format (Shamon et al., 2019). In line with this, Bansak and colleagues (2021) found that increasing the number of dimensions in the conjoint experiment table to as many as 18 only moderately influenced satisficing behavior, suggesting that the table format may indeed be easy for respondents to read. If the table format indeed has these intended effects, then respondents should be less likely to satisfice when reading the table presentation format than the text format, which generally uses more syllables and words, and longer sentences perhaps include more complex syntax, without further aiding the information intake. Based on satisficing theory, and replicating the predictions made in Shamon et al. (2019), the following hypotheses regarding respondent behavior will be assessed:

Hypothesis 1b (H1b): Respondents who evaluate the scenarios presented in a table format may produce data of greater quality (i.e., fewer refusals, fewer break-offs, fewer scenario non-responses, fewer total non-responses, and less dimension reduction) than respondents who evaluate the scenarios presented in a text format.

Hypothesis 1c (H1c): The effect that the presentation format may have on the total loss of information (in terms of total non-response) may be stronger for

respondents with lower educational attainment than for respondents with higher educational attainment.

Hypothesis 1d (H1d): The effect that the presentation format may have on the total loss of information (in terms of total non-response) may be stronger for older respondents than for younger respondents.

Paired Conjoint Experiment

In paired conjoint experiments, given that more information has to be processed and that the dimensions presented in the rows and columns of a table have to be correctly attributed to the correct set/profile, the table format may not outperform the text format to the same extent as in single conjoint experiments. The cognitive process may, therefore, change in the more complex paired setting, which, consequently, changes how the presentation format affects data quality and respondent experience. However, in line with the findings that the table format did outperform the text format in Hainmueller, Hopkins, and Yamamoto (2014), we still hypothesize that table format may outperform text format while remaining open to the reported differences between the single and paired conjoint experiment.

Hypothesis 2a (H2a): Respondents who evaluate scenarios presented in a table format may report a better respondent experience (i.e., report greater satisfaction with the questionnaire, take less time to evaluate the scenarios, and be more likely to participate in future panel waves) than respondents who evaluate the scenarios presented in text format.

Hypothesis 2b (H2b): Respondents who evaluate scenarios presented in a table format may produce data of greater quality (i.e., fewer refusals, fewer break-offs, fewer scenario non-responses, fewer total non-responses, and less dimension reduction) than respondents who evaluate the scenarios presented in a text format.

Hypothesis 2c (H2c): The effect that the presentation format may have on the total loss of information (in terms of total non-response rate) may be stronger for respondents with less educational attainment than for respondents with more educational attainment.

Hypothesis 2d (H2d): The effect that the presentation format may have on the total loss of information (in terms of total non-response rate) may be stronger for older respondents than for younger respondents.

Evaluation Criteria

To evaluate the hypotheses, in the present paper, the impact of the presentation format was categorized into aspects related to the respondent experience and the data quality. The same evaluation criteria were used to investigate the impact of presentation format in the single and the paired conjoint experiment.

The impact of the presentation format on respondent experience was investigated by assessing the cost of administration (in terms of processing time), the perceived experience of the survey (survey evaluation), and the probability of participation in subsequent waves of the Swedish Citizen Panel.

The impact of presentation format on data quality was investigated by assessing the refusal to participate in the survey experiment, the probability of the respondent breaking-off from completing the questionnaire, the probability of unanswered scenario evaluations, the number of faded-out dimensions, size of coefficient for dimensions, and respondents' response inconsistency. However, some respondents may very well have valid non-attitudes, meaning that a non-response, refusal, or break-off would be the most accurate representation of their evaluation. But, in line with the argument provided by Shamon et al. (2019), omitting to make a judgment will in the present paper be perceived to be an indication of satisficing and not as a valid representation of non-attitudes. The different answer behaviors refusal, break-off, non-response, and total non-response indicate that a respondent has applied a satisficing strategy, resulting in reduced data quality. Similarly, the varying importance that the respondent assigns to different dimensions is also proposed to be a form of satisficing, as the respondent reduces the cognitive burden of completing the questionnaire by excluding dimensions or assigning varying importance to the different dimensions in the experiment. Each of these forms of satisficing will be presented in more detail below, together with how each of them is operationalized.

Respondent Experience

Cost of Administration

The impact of the presentation format was investigated in terms of the time it took the respondents to answer the scenarios (i.e., the cost of administration). Longer administration times may be an indication that the respondents are struggling with interpreting and reading the vignette. In contrast, longer administration times may also be an indication that the respondents are paying attention to the information on the screen, leading to more thoughtful responses and greater data quality. Regardless of potential benefits to data quality, longer administration times will mean less time to ask other questions, as well as having to offer higher incentives to the respondent.

Time spent on the pages with the scenarios was used to assess the cost of administration. Due to an oversight in the survey programming, the time spent on the last scenario was not recorded for the single conjoint groups. Therefore, the cost of administration analyses for the single conjoint groups includes only the time spent on the first scenarios.¹ For the paired conjoint groups, the time spent on all scenarios was recorded and analyzed.

To reduce the impact of outliers, following Tukey (1977), total response times for the scenarios that were shorter than the interquartile range (IQR) of the sample response times * 1.5, and longer than the IQR * 1.5, were excluded from the cost of administration analysis. For the scenarios in the single conjoint groups, the lower bound for the excluded outliers was 0 seconds, and the upper bound was 285.9 seconds. For the scenarios in the paired conjoint groups, the lower bound was 0 and the upper bound was 520.3 seconds.

Survey Evaluation

The impact of the presentation format was also investigated in terms of survey evaluation. Respondents reported how well designed and how difficult the questionnaire was and rated their level of annoyance and concentration while filling out the questionnaire. A more positive overall survey evaluation may be an indication that the respondent found interpreting and evaluating the scenarios less challenging, and a more positive respondent experience may lead to better data quality.

Responses to the four survey evaluation questions (well-designed, difficult, annoyed, and needed concentration) were averaged into an index and coded to range from 0 to 1, with higher values indicating a more positive overall evaluation of the questionnaire.

Participation in Subsequent Panel Waves

Taking advantage of the ability to follow each respondent's participation in the Swedish Citizen Panel, the impact that the presentation format had on participation in the subsequent waves of the panel was investigated. Panelists were randomly sampled to be invited to complete studies in the subsequent waves of the Swedish Citizen Panel which led to that not all participating respondents in this study were invited to the subsequent waves of the panel. However, the majority of the panelists were invited. Participation/non-participation in subsequent waves may have many different causes, but a between-subject comparison of the respondents who saw the text and the respondents who saw the table format may reveal whether one of the formats was particularly detrimental to respondents' willingness to complete similar future tasks and experiments. A larger drop-out may be of particular interest for

1 The respondents were asked to evaluate four scenarios in total.

any sample provider attempting to estimate future costs of administering conjoint experiments.

Data Quality

Refusal

Refusing to evaluate all of the scenarios was one form of satisficing investigated. According to Shamon et al. (2019), respondents who refuse to respond decline to make judgments and thereby engage in a satisficing strategy by skipping at least one cognitive step when evaluating the scenarios. Refusal to make valid evaluations of all scenarios may indicate that the respondent found it more challenging to read or interpret the text or the table format, which has a clear negative impact on data quality.

Respondents were categorized as “refusals” if they did not evaluate any of the scenarios, answered “don’t know” in all scenarios, or provided no variation in their answers. Respondents who used these answering behaviors across all scenarios were coded as 1, and 0 otherwise.

Break-offs

Another form of satisficing strategy investigated was when respondents switched to constant non-valid answering behavior at some point after evaluating the first scenario. Opting for a non-valid answering behavior after having provided valid evaluations may indicate that the respondent found it more challenging to read or interpret the vignette and reduces the data quality.

Respondents were coded as 1 if they evaluated at least the first scenario and thereafter consistently used a non-valid answering strategy. That is, they were coded as 1 if they gave a valid answer to the first scenario and then started to answer “don’t know” or left a scenario evaluation unanswered, and 0 otherwise.²

Non-response

An alternative strategy for a respondent to decrease the cognitive burden of completing the questionnaire would be to alternate between validly evaluating scenarios and not validly evaluating scenarios (Shamon et al., 2019). Such a strategy should be considered a weaker form of satisficing than refusal or breaking-off but remains a negative influence on data quality. This evaluation criterion aims to capture the type of satisficing behavior where the respondent remains in the experiment (hence, does not refuse to answer, or break off answering the questions) but instead alternates between validly judging and not validly judging scenarios to make the survey

2 The respondents were asked to evaluate four scenarios in total.

easier to complete. Fewer invalidly judged scenarios indicate that the respondent found interpreting and reading the text or the table format less challenging and suggest better data quality.

Scenarios evaluated by respondents (which were not coded as refusal or break-off) were coded as a non-response if the respondents invalidly judged at least one scenario but not all of them. A scenario was invalidly judged if the respondent answered “don’t know” or did not provide an answer. Note that this evaluation criterion was computed at the scenario level and not at the respondent level.

Total Non-response

To capture the total loss of information due to the presentation format, the total non-response was computed and captured all the scenarios that were invalidly judged, irrespective of the type of strategy used by the respondent. The criteria were computed on the scenario level, and a scenario was coded as 1 (total non-response) if it was invalidly judged by either not answering or answering “don’t know,” or if the respondent provided no variation in the answer across all four scenarios or broke-off their participation, and 0 otherwise.

Response Inconsistency and Partial Dimension Reduction

A presentation format that respondents have a harder time reading or understanding, or that makes it more difficult for respondents to distinguish between the dimensions may produce a weaker predictive ability of the attribute levels on the dependent variable of the conjoint experiment. As a result of increased response inconsistency, an underperforming presentation format may produce more measurement errors in the dimensions’ predictions. Hence, a presentation format that yields the largest estimated parameters for the dimensions and has the lowest measurement error should be interpreted as the more valid and preferable format to use for conjoint experiments.

Furthermore, a cognitively more burdensome presentation format should have a stronger detrimental effect on both parameter estimates and measurement error as a respondent evaluates more scenarios (i.e., a partial dimension reduction). A more burdensome presentation format should increase the likelihood of the respondent putting less and less cognitive effort into distinguishing between different dimensions as the number of evaluated scenarios increases. Hence, one would expect to see a weaker and weaker predictive ability of the dimensions on the dependent variables as well as greater measurement error across scenarios (i.e., one would expect to see a reduction in the impact that the dimensions have on the dependent variables).

To investigate the partial dimension reduction and response inconsistency, the invariance in parameters and the invariance in error variance were compared across the two presentation formats by applying the structural equation modeling (SEM)

technique to predict/make a judgment on salary or party preference of respondents based on the dimensions presented in the conjoint experiment (MacDonald, 2016). All exogenous predictors were free and allowed to covary.³

Methods and Materials

Sample

The respondents were a pre-stratified sample of members of the Swedish Citizen Panel run by the Laboratory of Opinion Research (LORE) at the University of Gothenburg, Gothenburg, Sweden. At the time of the study, the Swedish Citizen Panel consisted of about 59,000 self-selected panelists, and members of the panel were invited to complete approximately four online omnibus questionnaires each year. The panelists were, therefore, relatively experienced, and were not paid an incentive to complete the questionnaires.

The presentation format experiment was administered to 7,000 panelists pre-stratified by sex (male, female), age (18–34, 35–49, 50–85 years), and education (low/middle education: less than 3 years of post-secondary education, high education: 3 or more years of post-secondary education) between February 24th, 2020, and March 19th, 2020. For the demographic distributions, see Table 1. Reminders were sent on March 3rd, 2020, and on March 11th, 2020, to all respondents who had not yet completed the questionnaire. Out of the 7,000 respondents invited to participate, 4,236 completed the experiment (American Association for Public Opinion Research (AAPOR) response rate 5 (RR5): 59%).

3 Parameters were estimated using the function SEM in Stata 16, with the group option and all other options set at default. By default, SEM in Stata 16 allows all exogenous predictors to covary.

Table 1 Demographic distributions of the experiment sample and the Swedish population, and the difference between the sample and Swedish population.

Variables	Sample	Population	Difference
Age, years			
18–34	18%	28%	-10%
35–49	23%	25%	-2%
50–85	59%	47%	+12%
Gender			
Male	48%	50%	-2%
Female	52%	50%	+2%
Education			
Low/middle	68%	76%	-8%
High	32%	24%	+8%

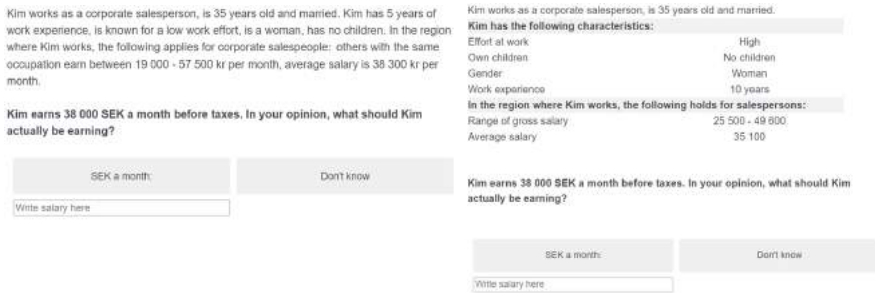
Notes. N=4,236.

Procedure

Each respondent was randomly assigned to either a single or paired conjoint experiment, and within each experiment, respondents were randomly assigned to see the conjoint in either a text or a table presentation format.

Single Conjoint Experiment

Respondents assigned to the single conjoint experiment reported the amount of salary a person deserved to earn, based on four dimensions of the person (sex, number of children, work experience, and work effort) and two contextual dimensions (average salary for others in the same region and range of salaries for people of the same occupation), approximating a direct replication of the experiment in Shamon et al. (2019). See Table S1 in the Supplementary Online Materials (SOMs) for the dimension levels. Whether the person's dimensions or the contextual dimensions were presented first was randomly determined for each participant. In addition, each level of the dimensions was determined randomly. Each respondent was presented with four scenarios to evaluate. See Figure 1 for a screenshot of the single conjoint experiment translated to English and SOM S1.1. for the full questionnaire logic.



Notes. Respondents were randomized to either the single or the paired conjoint experiment. Respondents in the single conjoint experiment group were randomized to read the information in either text or table presentation format.

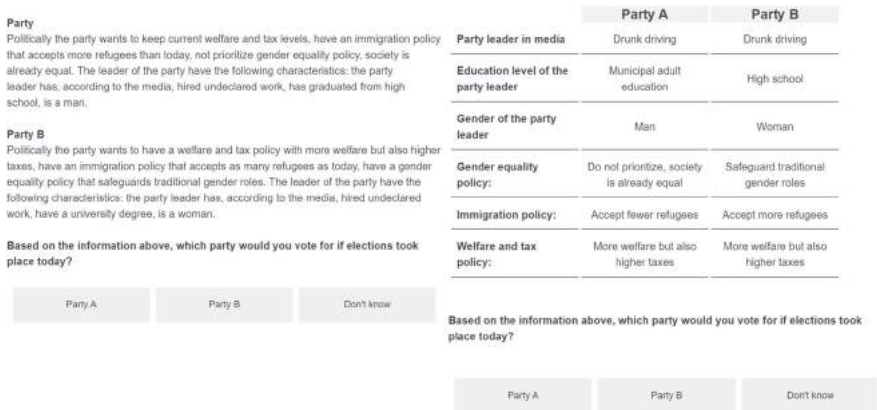
Figure 1 Screenshot of the single conjoint experiment administered on computers.

Paired Conjoint Experiment

Respondents assigned to the paired conjoint experiment reported which of two hypothetical political parties they would vote for and how likely they would be to vote for each of the parties. The paired conjoint experiment did not use the above topic (i.e., the salary of a worker) because the single conjoint experiment did not lend itself to be easily translated into a paired experiment. Instead, the paired experiment was developed to closer mimic Hainmueller et al.’s (2015) experiment where respondents evaluated two political agents (in our case, political parties). The two parties were described in terms of three political dimensions (immigration, welfare and taxes, and equality policy stances) and three dimensions of their respective party leader (sex, educational attainment, and media image). See Table S2 in the SOM for the dimension levels. Whether the leader’s or the party’s dimensions were presented first was randomly determined for each respondent. In addition, each level of the dimensions was determined randomly.

Each respondent was presented with four scenarios to evaluate. See Figure 2 for a screenshot of the paired conjoint experiment translated to English and SOM S1.4 for the full questionnaire logic.

After the presentation format experiment, the respondents reported their level of annoyance and concentration while filling out the questionnaire, as well as how well-designed and difficult the questionnaire was.



Notes. Respondents were randomized to either the single or the paired conjoint experiment. Respondents in the paired conjoint experiment group were randomized to read the information in either text or table presentation format.

Figure 2 Screenshot of the paired conjoint experiment administered on computers.

Differences from Shamon et al. (2019)

The procedure and sample in this paper diverge from a direct replication of Shamon et al. (2019) in three ways: Firstly, Shamon et al. administered a third condition of the text format; underlining the dimensions in the text vignette. We did not implement that condition.⁴

Secondly, Shamon et al. (2019) presented dimensions in a fixed order, whereas we randomly assigned whether the dimensions of the person or dimensions of the context were presented first for each respondent (and similarly for the order of the personal/party leader and contextual/party dimensions). The order was randomized to reduce recency and primacy effects, as well as to avoid the order effect cautioned

⁴ The ease of reading a paragraph (i.e., processing fluency) has been found to correlate with both actual cognitive effort and perceived effort (Reber, Schwarz, and Winkielman, 2004; Song and Schwarz, 2008). That is, a paragraph containing cursive/italized letters has been found to be more difficult to process than a simple font such as Arial (Song and Schwarz, 2008). Similarly, underlining certain phrases likely presents respondents with yet another layer of cognitive burden compared to an easy-to-read paragraph with less clutter. This notion is supported in Shamon et al.'s (2019) findings, where the underlined text format took respondents longer to process than the other formats. Furthermore, underlining has to be accurately understood by each respondent as “the important information” in order to actually improve data quality. If the respondents interpret the underlining differently, the result may be more random measurement error.

by Auspurg and Hinz (2015) (see also Auspurg & Jäckle, 2017) but disregarded in Shamon et al. (2019).

Thirdly, Shamon et al. (2019) presented respondents with 16 scenarios that they had to evaluate, whereas we asked the respondents to evaluate four scenarios. Respondents were presented with fewer scenarios to resemble the questionnaires that they usually complete and to lower the risk of exhausting respondents. Administering fewer scenarios may contribute to weaker effects on outcomes that correlate with questionnaire fatigue, but we opted for more unique observations over having more scenarios in order to increase the variation of the type of respondents. Therefore, instead of presenting respondents with many scenarios, we included a larger sample of respondents ($N=2,068$ in the single conjoint experiment) compared to Shamon et al. ($N=498$), thus following Shamon et al.'s (2019, p. 34) suggestion to increase statistical power in order to be able to identify small effects.

However, in line with Shamon et al.'s (2019) approach, we decided to still treat respondents' answers as refusals if they provided the exact same answers for all four scenarios. Although the number of scenarios was fewer, we deem it unlikely that a respondent would validly consider all dimensions and still provide the same salary four times in a row, or in the paired scenario, always choosing Party A or Party B. Lastly, our sample included respondents older than 69 years in an attempt to better generalize to the general population compared to Shamon et al. (2019).

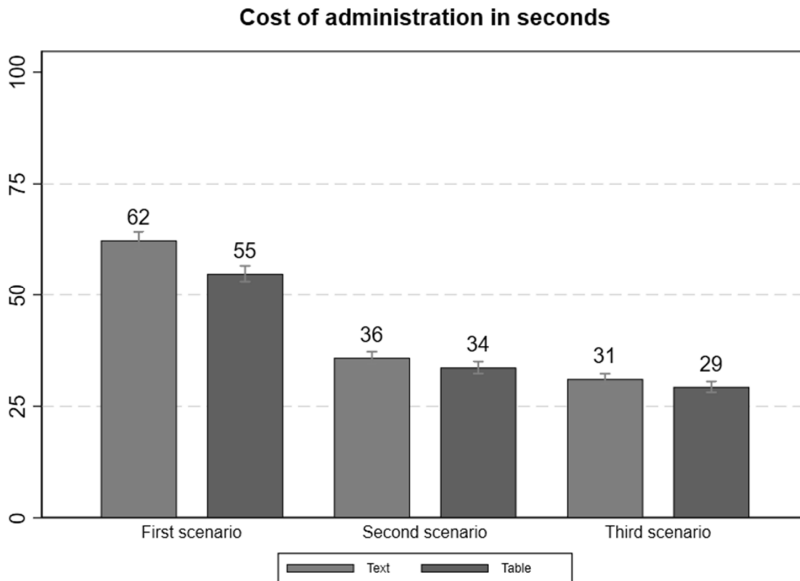
Results

In this section, the impact that the presentation format had on the single conjoint experiment will be presented first, followed by the impact it had on the paired conjoint experiment. The evaluation of the presentation format will be separated into aspects related to the participant experience (cost of administration, evaluation, participation in subsequent waves) and aspects related to data quality (refusal, break-off, non-response, total non-response, dimension reduction, and moderation of effects).

Single Conjoint Experiment

Respondent Experience

Cost of administration. Across the first three scenarios, respondents who made a judgment on the salary of the worker when reading about the dimensions in a text paragraph took statistically significantly 12 seconds longer to submit their evaluations ($M = 129$ seconds, standard deviation (SD) = 54) than the respondents who evaluated the salary when the dimensions were presented in a table format ($M =$



Notes. N=1,874. Respondents who answered the three scenarios for which time was recorded, and whose response times were not longer than 1.5 times the interquartile range (IQR) for the three scenarios, were included in the analyses (N excluded = 137).

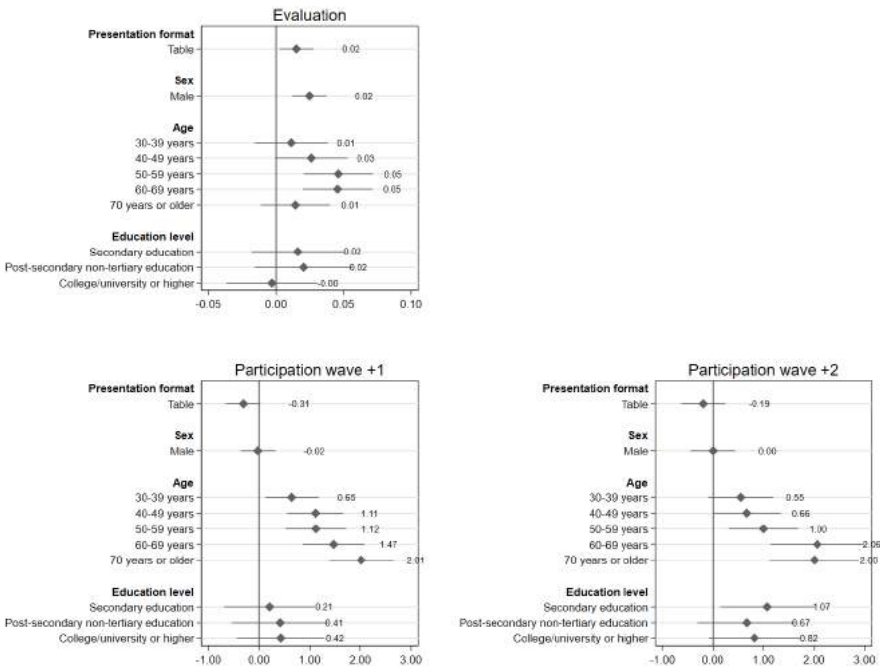
Figure 3 Cost of administration, in seconds, for the single conjoint experiment.

117 seconds, SD = 53; $b = -11.72$, standard error (SE) = 2.47, $p < 0.001$), a statistically significant difference providing support for H1a.

However, the difference in administration time was greatest for the first scenario (see Figure 3). In the second and third scenario, the difference in administration time between the text and the table version statistically significantly decreased ($b_{table * scenario 2} = 5.22$, SE=1.29, $p < 0.001$; $b_{table * scenario 3} = 5.68$, SE=1.30, $p < 0.001$) (see SOM S2.1, Table S3).

Over scenarios, respondents were able to reduce the time to evaluate the specific scenario but the respondents who read the text format reduced their processing time more compared to those who read the table format (see Figure 3). However, as will be shown in the dimension reduction analyses, the stronger reduction in processing time over scenarios for respondents presented with text format seems to have stemmed from the fact that those respondents invested less and less cognitive effort in their evaluations (see section *Moderation effects*).

Evaluation. Respondents who read the table presentation format reported a more positive evaluation of the questionnaire ($b=0.02$, SE=0.01, $p < 0.01$) than the respondents who read the text presentation format (see Figure 4). The significant



Notes. N=1,968 (Evaluation); N=1,484 (Participation wave +1); N=903 (Participation wave +2). The number of observations differs in the three panel waves because panelists were randomly sampled to be invited to complete the panel wave or not. Regression coefficients (gray diamonds) from one ordinary least squares regression (OLS) and two logistical regressions with their respective 95% confidence intervals (CIs) (gray solid lines). A positive value of the coefficient indicates a higher overall evaluation or a higher likelihood of participation in subsequent waves. Baseline categories were female, 18–29 years of age, and compulsory education (9 years).

Figure 4 Respondent experience in terms of overall questionnaire evaluation and participation in the subsequent waves, for the single conjoint experiment.

effect of presentation format was found in both bivariate analyses and when including controls (see Figure 4 and SOM S2.1., Table S4). Male respondents were, overall, more positive than female respondents ($b=0.02$, $SE=0.01$, $p<0.001$), and older respondents, in the 50–59 and 60–69-year groups, were more positive than those aged 29 years or younger (see Figure 4).

When analyzing the four separate evaluation questions used to construct the index of overall survey evaluation, the only significant effect of presentation format was found for the question asking how annoyed the respondent was when filling out the questionnaire ($b=0.04$, $SE=0.01$, $p<0.01$; see SOM S2.1., Table S5).

Participation in subsequent waves of the Swedish Citizen Panel. Presentation format had an initial significant effect on participation in the immediate subsequent wave of the Swedish Citizen Panel, but this effect disappeared with later waves (see Figure 4).

Respondents who were given the table presentation format were marginally less likely to participate in the following wave of the Swedish Citizen Panel, which was administered approximately 3 months after the presentation format experiment (wave +1) ($b=-0.31$, $SE=0.17$, $p<0.10$). However, the effect of presentation format was not statistically significant in the wave of the Swedish Citizen Panel distributed approximately 6 months after the presentation format experiment (wave +2) (see Figure 4). The age of the respondent had a significant effect on participation, both in the first and second wave following the presentation format experiment, with older respondents being more likely to participate in subsequent waves. Participation in the second wave following the experiment was significantly more likely among respondents with upper secondary education ($b=1.07$, $SE=0.47$, $p<0.01$).

However, on average, the results provide additional support for H1a, as respondents reported both greater satisfaction and took less time to evaluate the scenarios when receiving the table format compared to the text format. The results on participation in the subsequent waves confirm this and the immediate negative effect of the table presentation format disappeared after the first wave following the presentation format experiment.

Data Quality

Table 2 presents a descriptive summary of the sample size and answer behaviors of respondents assigned to the single conjoint experiment. Overall, 482 respondents presented with the single conjoint experiment chose to either break-off or refusal to answer the scenarios (see Table 2). Similar patterns were found between respondents presented with the text and the table presentation format. However, 26 (2.5%) respondents presented with the table format chose to stop filling out the questionnaire compared to 16 (1.5%) of those who saw the text presentation format (Table 2). The most commonly used satisficing answering behavior on the respondent level was to provide no answers or don't know answers across all of the four scenarios, 167 (16%) respondents in the text format and 161 (15.8%) respondents in the table format.

Table 3 presents a similar descriptive summary of the answering behavior at the scenario level, and the most common satisficing answering behavior was a refusal to answer any of the scenarios. The text and table presentation format yielded roughly the same amount of total loss of information (total non-response) (text: 24.4%, table: 25%) but the text presentation format had slightly fewer respondents breaking-off (1.5%) compared to the table presentation format (2.5%).

Table 2 Sample sizes and answer behavior at the respondent level, for the single conjoint experiment.

Experiment setting	Sample size	No evaluated scenarios (refusal, don't know) (1)	Invalid constant answer behavior (2)	Refusals (1) + (2)	Break-off (3)	Total (1) + (2) + (3)
Text	1047	167 (16%)	57 (5.4%)	224 (21.4%)	16 (1.5%)	240 (22.9%)
Table	1021	161 (15.8%)	55 (5.4%)	216 (21.2%)	26 (2.5%)	242 (23.7%)
Sum	2068	328 (15.9%)	112 (5.4%)	440 (21.3%)	42 (2%)	482 (23.3%)

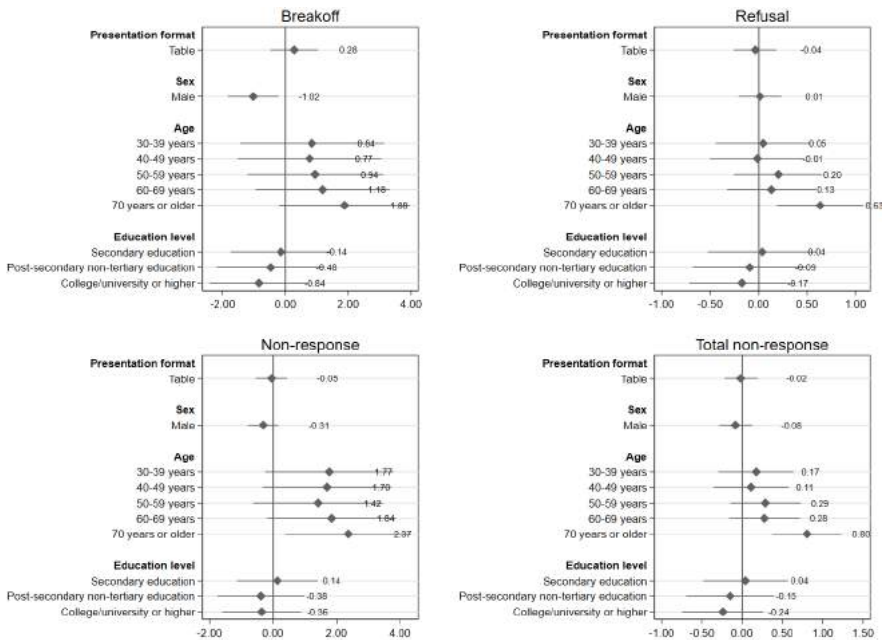
Notes. Results at the respondent level.

Table 3 Sample sizes and answer behavior at scenario level, for the single conjoint experiment.

Experiment setting	Gross sample size (N = 2,068)			Net sample size after excluding refusals and break-offs (N = 1,586)	
	Gross sample of scenarios	Refusals (1) + (2)	Break-off (3)	Non-response (4)	Total non-response (1) + (2) + (3) + (4)
Text	4,188	896 (21.4%)	64 (1.5%)	63 (1.5%)	1,023 (24.4%)
Table	4,084	864 (21.2%)	104 (2.5%)	55 (1.3%)	1,023 (25%)
Sum	8,272	1760 (21.3%)	168 (2%)	197 (2.4%)	2,046 (24.7%)

Notes. Results at scenario level. Gross sample of scenarios is calculated by multiplying the group size of an experimental setting by the set size (4 scenarios per respondent).

Refusals, break-offs, non-responses, and total non-response. The two presentation formats did not differ with regard to refusals, break-offs, non-responses, and total non-response (see Figure 5). Although the presentation format statistically significantly affected non-response (respondents who altered between judging and not judging a vignette, excluding refusals and break-offs), it did so only for the first scenario, where the text format resulted in greater levels of non-response. The effect then disappeared in the subsequent three scenarios, and on average, there was no effect of presentation format on non-responses, nor on refusals, break-offs, and total non-response. The results were not moderated by how many scenarios the respondents had answered.



Notes. N=1,480 (Break-off); N=1,965 (Refusal); N=6,067 (Non-response); N=7,866 (Total non-response). Regression coefficients (gray diamonds) from four logistical regressions with their respective 95% confidence intervals (CIs) (gray solid lines). A positive value indicates a higher likelihood of break-off, refusal, non-response, or total non-response. Baseline categories were female, 18–29 years of age, and compulsory education (9 years). Results on break-offs and refusals are at the respondent level, while results on non-response and total non-response are at the scenario level. We controlled within-participant clustering for non-response and total non-response using cluster-robust standard errors.

Figure 5 Data quality in terms of break-offs, refusals, non-responses, and total non-response, for the single conjoint experiment.

Respondents 70 years or older were more prone to refusal ($b=0.63$, $SE=0.23$, $p<0.01$), non-response ($b=2.37$, $SE=1.01$, $p<0.05$), and total non-response ($b=0.80$, $SE=0.22$, $p<0.01$) compared to the baseline (18–29 years old). Furthermore, women were more likely to breaking-off ($b=-1.02$, $SE=0.41$, $p<0.05$) than men (see SOM S2.1., Table S8).

Moderation effects. To test whether the effect of presentation format on data quality was moderated by education (H1c) and age (H1d), two new models were estimated predicting total non-response with presentation format, age, education, gender, and with an interaction between either presentation format and age or presentation format and education. All graphs on moderating effects can be found in SOM S2.1., Figures S1–S4.

Education. In contrast to the expected, education did not moderate the effect that the presentation format had on the probability of total non-response. Respondents with lower educational attainment were not more likely to adopt an invalid answer behavior due to the presentation format. Hypothesis 1c was, therefore, not supported.

Age. Similar to the effect found for educational attainment, age did not moderate the effect that the presentation format had on total non-response. Older respondents were more likely to adopt an invalid answer behavior, but there was no significant difference in the effect that the presentation format had by age of the respondents. Therefore, although older respondents found it more demanding to fill out the questionnaire, they did not find a certain presentation format more demanding compared to younger respondents, providing no support for H1d.

Response inconsistency and partial dimension reduction. In contrast to the hypothesis (H1b), in the first scenario, respondents who were presented with the table presentation format yielded statistically significantly greater response inconsistency (i.e., weaker prediction of a dimension) ($b_{\text{high effort dimension}} = -2,117$, $SE = 522$, $\chi^2(1, 1,435) = 16.76$, $p < 0.001$), albeit not with significantly more measurement error ($\epsilon = 24,141,723$, $SE = 1,206,417$), compared to those who were given the text format ($\epsilon = 22,937,796$, $SE = 1,279,509$, $\chi^2(1, 1,435) = 0.47$, $p = 0.49$) (see Table 4, column 1). This underperformance in prediction strength remained for scenario 2 and 3 (see Table 4, column 2 and 3). However, by the fourth scenario, the performance difference had shifted to the table format yielding statistically significantly stronger predictions for three of the dimensions and produced statistically significantly less measurement error of the prediction ($\epsilon_{\text{difference table versus text}} = -16,657,746$, $\chi^2(1, 1,435) = 31.77$, $p < 0.001$) (see Table 4, column 4).

The reversal in the outcome, from the text presentation format outperforming the table format in the first scenarios to the table format heavily outperforming the text format in the last scenario, is evidence that the text format suffered from a stronger dimension reduction than the table format. This increased partial dimension reduction across scenarios for the text format is well-illustrated by the much faster increased measurement error across scenarios among those receiving the text format compared to those receiving the table format. Respondents reading the text format evaluated the attributes with increasing measurement error over scenarios ($\epsilon_{\text{scenario 2} - \text{scenario 1}} = 4,390,654$, $\chi^2 = 5.47$, $p < 0.05$; $\epsilon_{\text{scenario 3} - \text{scenario 2}} = 5,815,856$, $\chi^2 = 6.63$, $p < 0.01$; $\epsilon_{\text{scenario 4} - \text{scenario 3}} = 13,945,123$, $\chi^2 = 21.20$, $p < 0.01$), whereas respondents who saw the table format remained consistent in the amount of measurement error they produced ($\epsilon_{\text{scenario 2} - \text{scenario 1}} = 2,277,566$, $\chi^2 = 1.44$, $p = 0.23$; $\epsilon_{\text{scenario 3} - \text{scenario 2}} = 4,310,024$, $\chi^2 = 4.03$, $p < 0.05$; $\epsilon_{\text{scenario 4} - \text{scenario 3}} = -297,624$, $\chi^2 = 0.02$, $p = 0.90$). Furthermore, this shift occurred even though respondents who resorted to satisficing behavior through refusal, breaking-off, or not responding were already excluded from the dimension reduction analyses. Consequently, even

Table 4 Parameter differences between text and table format predicting salary with the dimensions, for the single conjoint experiment

	Parameter differences (table – text)			
	Scenario 1	Scenario 2	Scenario 3	Scenario 4
Dimensions				
Female	-798 (516)	-754 (550)	-1,148 ⁺ (599)	694 (668)
Two children	601 (521)	-183 (552)	226 (616)	389 (669)
10 years of experience	213 (518)	-81 (552)	-1,311* (606)	2,001** (661)
High effort	-2,117*** (522)	-1,918*** (554)	-1,547* (609)	6,611*** (667)
Medium salary	-180 (605)	434 (666)	421 (750)	-460 (785)
High salary	-610 (661)	459 (689)	328 (723)	2,596** (845)
Others earn 25,500–49,600 SEK	-707 (513)	927 ⁺ (553)	673 (623)	1,636* (671)
Constant	29,383*** (533)	30,776*** (558)	30,149*** (651)	37,554*** (756)
Error variance, table	24,141,723*** (1,279,509)	26,419,283*** (1,400,219)	30,729,307*** (1,628,650)	30,431,683*** (1,612,876)
Error variance, text	22,937,796*** (1,206,417)	27,328,450*** (1,437,344)	33,144,306*** (1,743,230)	47,089,429*** (2,476,676)
Error variance difference (table – text)	1,203,927	-909,167	-2,414,999	-16,657,746***
χ^2 of difference	0.47	0.21	1.03	31.77
Observations	1,435	1,435	1,435	1,435

Notes. Regression coefficients from four ordinary least squares (OLS) regression equations, standard errors (SEs) in parentheses. Positive parameters mean that the table format outperformed the text format in predicting the person's salary, whereas negative parameters mean that the text outperformed the table format. Omitted dimensions were "no children," "5 years of experience," "low effort," "low salary," "others earn 19,000–57,500 SEK." Only the respondents who answered all four scenarios were included. See SOM S2.1., Table S9, for the parameters, separately for presentation format and scenario.

⁺p<0.1, *p<0.05, **p<0.01, ***p<0.001.

though the text format outperformed the table format in the first scenario, the win was short-lived and, overall, the table format produced less dimension reduction.

Hence, for the single conjoint experiment, there was no clear support for the hypothesis that the table format would outperform the text format in terms of data quality (H1b). With regard to refusals, break-offs, non-responses, and total non-response, there was no support for that hypothesis. For partial dimension reduction, the results depend on the number of scenarios the researcher wishes to include: if the respondents evaluate one scenario, the data favor text format, but when evaluating more scenarios, table format would be preferred.

Paired Conjoint Experiment

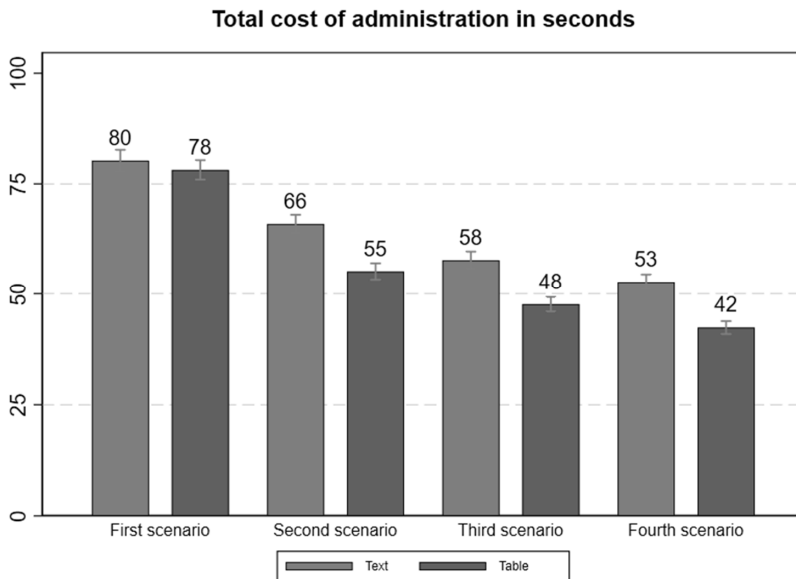
Respondent Experience

Cost of administration. Across the four scenarios, the respondents who evaluated the two political parties after reading the text format took, as hypothesized in H2a, statistically significantly 32 seconds longer to submit their evaluations ($M=255$ seconds, $SD=94$) than the respondents who evaluated the parties using the table format ($M=223$ seconds, $SD=89$; $b=-32.19$, $SE=4.06$, $p<0.001$).

In contrast to the single conjoint experiment, the difference in administration time between the text and the table version was the smallest for the first scenario. The differences between the formats subsequently increased as the respondents evaluated more scenarios (see Figure 6). In the second, third, and fourth scenario, the differences in administration time between the text and the table formats were statistically significantly shorter for the table format than for the text format ($b_{\text{table}} * \text{scenario 2} = -8.66$, $SE= 1.69$, $p<0.001$; $b_{\text{table}} * \text{scenario 3} = -7.98$, $SE=1.71$, $p<0.001$; $b_{\text{table}} * \text{scenario 4} = -8.31$, $SE=1.70$, $p<0.001$) (see SOM S2.3., Table S10). As will be shown in the analyses of partial dimension reduction (see section *Moderating effects*), in contrast to the single conjoint experiment, the shortening of processing time for the paired conjoint experiment did not correspond to a stronger dimension reduction. Therefore, the shorter time to process the table presentation format seems to have been preferable over the, in total, longer processing time of the text presentation format.

Evaluation and participation in subsequent waves of the Swedish Citizen Panel. In contrast to the single conjoint experiment, the presentation format in the paired conjoint experiment had no significant effect on overall survey evaluation or participation in subsequent waves.

Respondents who received the table format did not evaluate the questionnaire in a significantly more positive or negative way; nor did they evaluate the separate evaluation question significantly differently than the respondents who read the text format. Male respondents evaluated the questionnaire in a significantly more positive way than female respondents ($b=0.03$, $SE=0.01$, $p<0.001$). In contrast to the



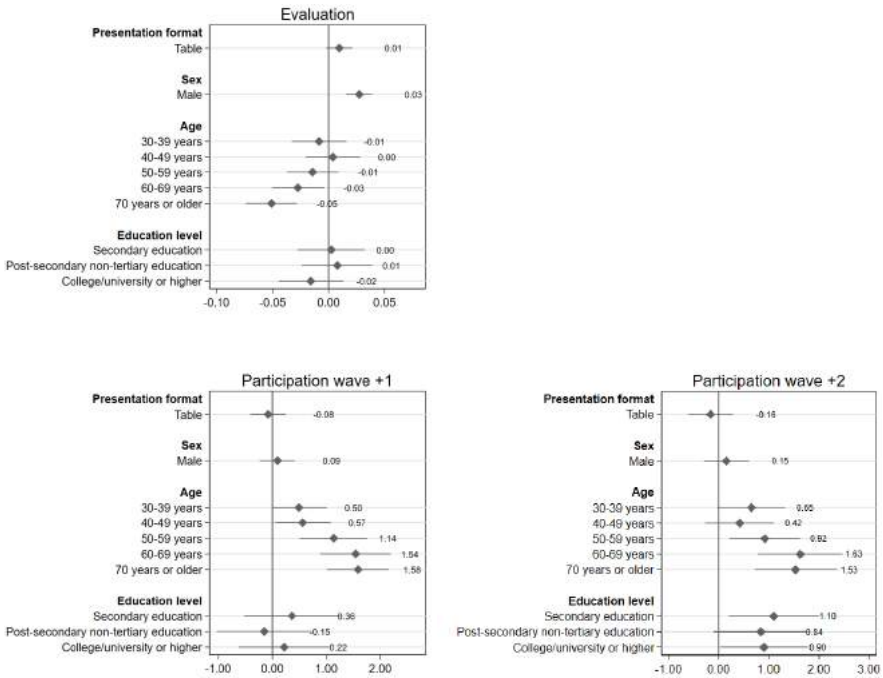
Notes. N=2,037. Respondents who answered the four scenarios and whose response times were not longer than 1.5 times the interquartile range (IQR) were included in the analyses (N excluded = 124).

Figure 6 Cost of administration, in seconds, for the paired conjoint experiment.

single conjoint experiment, in the paired conjoint experiment older respondents in the 60–69-year and over 70-year age groups gave an overall more negative survey evaluation compared to respondents under 29 years old (see Figure 7).

Respondents who read the table presentation format were not significantly more or less likely to participate in subsequent waves of the Swedish Citizen Panel following the presentation format experiment (see Figure 7). Again, the probability of participation in subsequent waves of the Swedish Citizen Panel was significantly higher among older respondents in the paired conjoint than in the single conjoint experiment. Participation in the second wave was also significantly more likely among respondents who had an upper secondary education ($b=1.10$, $SE=0.46$, $p<0.01$) or college/university education ($b=0.90$, $SE=0.44$, $p<0.01$) (see SOM S2.3., Table S12).

Therefore, the overall result provides only partial support for H2a. Respondents on average took less time to complete the survey when seeing the table compared to text, but there was no support for a difference in satisfaction with the survey or future participation depending on the presentation format.



Notes. N=2,125 (Evaluation); N=1,536 (Participation wave +1); N=965 (Participation wave +2). The number of observations differs in the three panel waves because panelists were randomly sampled to be invited to complete the panel wave or not. Regression coefficients (gray diamonds) from one ordinary least squares regression (OLS) and two logistical regressions with their respective 95% confidence intervals (CIs) (gray solid lines). A positive value indicated higher overall evaluation (Evaluation) or a higher likelihood of participation in subsequent waves (Participation wave +1 and Participation wave +2). Baseline categories were female, 18–29 years of age, and compulsory education (9 years).

Figure 7 Respondent experience in terms of overall questionnaire evaluation and participation in the subsequent waves, for the paired conjoint experiment.

Data Quality

Table 5 presents a descriptive summary of sample sizes and answer behaviors at the respondent level in the paired conjoint experiment. Descriptively, the text presentation format caused more respondents to break-off or refuse to evaluate the scenarios compared to the table presentation format, 292 (27.6%) respondents in the text format compared to 224 (19.9%) respondents in the table format (see Table 5).

The most common satisficing answering behavior for respondents presented with the text presentation format was to break-off. For respondents presented with the table presentation format, however, providing an invalid answering behavior,

such as providing no variation in their answers, was the most common satisficing answering behavior (see Table 5).

The results presented at the scenario level further show that the total loss of information (total non-response) seemed descriptively greater with the text presentation format compared to the table presentation format, 1,687 (39.8%) incorrectly judged scenarios in the text presentation format compared to 1,398 (31%) in the table format (see Table 6).

Refusals, break-offs, non-responses, and total non-response. When being presented with the table presentation format, respondents were statistically significantly less likely to adopt a refusal answering behavior, that is, to not answer, to constantly provide “don’t know” answers, or to not vary their responses across the four scenarios ($b=-0.29$, $SE=0.12$, $p<0.01$), compared to when receiving the text

Table 5 Sample sizes and answer behavior at the respondent level, for the paired conjoint experiment.

Experiment setting	Sample size	No evaluated scenarios (refusal, don't know) (1)	Invalid constant answer behavior (2)	Refusals (1) + (2)	Break-off (3)	Total (1) + (2) + (3)
Text	1,058	98 (9.2%)	72 (6.8%)	170 (16.1%)	122 (11.5%)	292 (27.6%)
Table	1,126	60 (5.3%)	84 (7.5%)	144 (12.8%)	80 (7.1%)	224 (19.9%)
Sum	2,159	158 (7.3%)	156 (7.2%)	314 (14.5%)	202 (9.4%)	516 (23.9%)

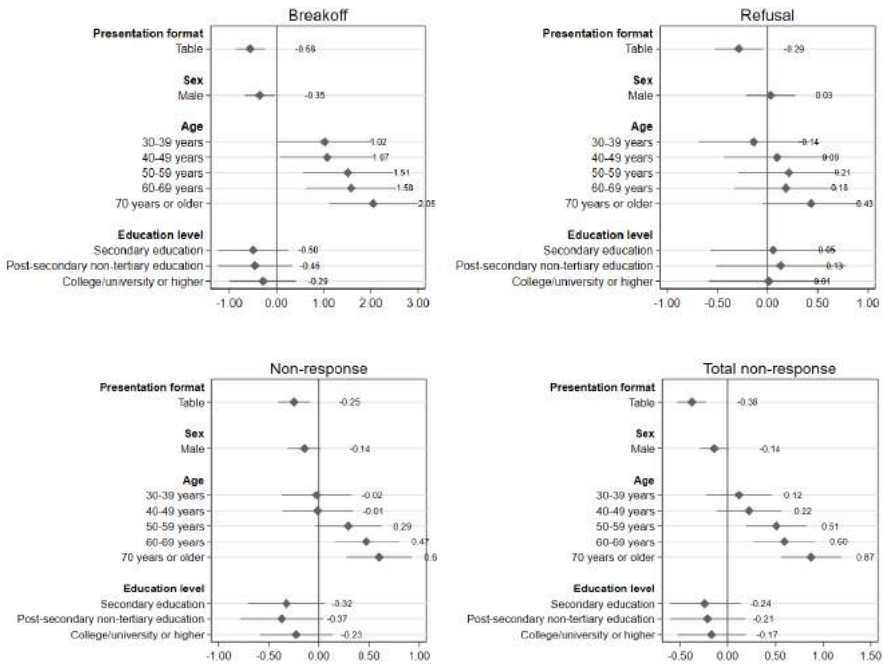
Notes. Results at the respondent level.

Table 6 Sample sizes and answer behavior at scenario level, for the paired conjoint experiment.

Experiment setting	Gross sample size (N = 2,159)			Net sample size after excluding refusals and break-offs (N = 1,643)	
	Gross sample of scenarios	Refusals (1) + (2)	Break-off (3)	Non-response (4)	Total non-response (1) + (2) + (3) + (4)
Text	4,232	680 (16.1%)	488 (11.5%)	519 (12.2%)	1,687 (39.8%)
Table	4,504	576 (12.8%)	320 (7.1%)	502 (11.1%)	1,398 (31%)
Sum	8,636	1,256 (14.5%)	808 (9.4%)	1021 (11.8%)	3,098 (35.8%)

Notes. Results at scenario level. Gross sample of scenarios is calculated by multiplying the group size of an experimental setting by the set size (4 scenarios per respondent).

presentation format. The significant effect of presentation format on refusals was found in both the bivariate analysis and when including controls (see Figure 8 and SOM S2.4., Tables S13 and S14). No other significant effects on refusals were found.



Notes. N=1,818 (Break-off); N=2,125 (Refusals); N=7,129 (Non-response); N=8,489 (Total non-response). Regression coefficients (gray diamonds) from four logistical regressions with their respective 95% confidence intervals (CIs) (gray solid lines). A positive value indicates a higher likelihood of break-off, refusal, non-response, or total non-response. Baseline categories are female, 18–29 years of age, and compulsory education (9 years). Results on break-offs and refusals are at the respondent level while results on non-response and total non-response are at the scenario level. We controlled within-participant clustering for non-response and total non-response using cluster-robust standard errors.

Figure 8 Data quality in terms of break-off, refusals, non-responses, and total non-response, for the paired conjoint experiment.

Similarly, respondents who saw the table presentation format were significantly less likely to start giving valid evaluations and then switch to a constant non-valid answer behavior (break-off) compared to respondents presented with the text format ($b=-0.56$, $SE=0.16$, $p<0.001$). Furthermore, male respondents were significantly less likely to breaking-off than women ($b=-0.35$, $SE=0.16$, $p<0.01$). The

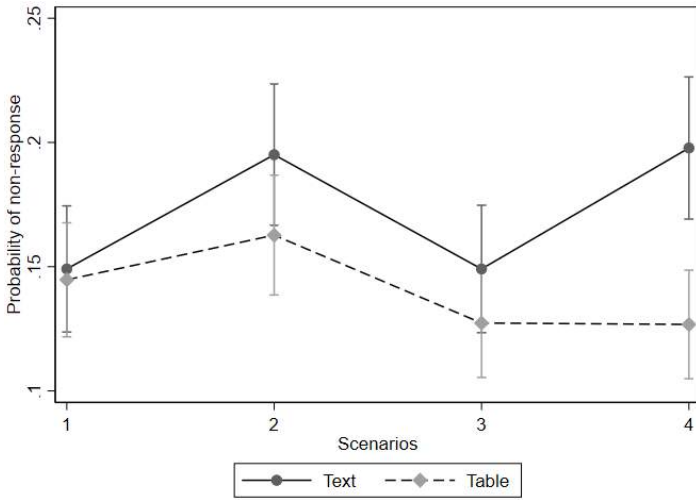
probability of a break-off increased with the age of the respondents in an almost linear fashion, where the older the respondent the higher the probability of breaking-off (see Figure 8).

The table presentation format also performed better with regard to non-response (i.e., alternating between validly judging and not validly judging scenarios). Respondents who saw the table format were less likely to resort to a non-response behavior ($b=-0.25$, $SE=0.08$, $p<0.01$) compared to those seeing the text format, when controlling for gender, age, and education, on non-response. Furthermore, the effect was moderated by the number of scenarios, where the adverse effect of using text format became evident only in the last scenario, where there were statistically significantly fewer non-responses among the table format respondents ($b_{\text{table} * \text{scenario } 4} = -0.50$, $SE=0.18$, $p<0.01$).

The predictive probability of a non-response in the fourth scenario was 0.20 for the text format and 0.13 for the table format (see Figure 9). Respondents who adopted a refusal or breaking-off behavior were excluded. In line with previous results, older respondents were more likely to exhibit a non-response behavior, that is, to switch/alternate between validly answering and not validly answering a scenario.

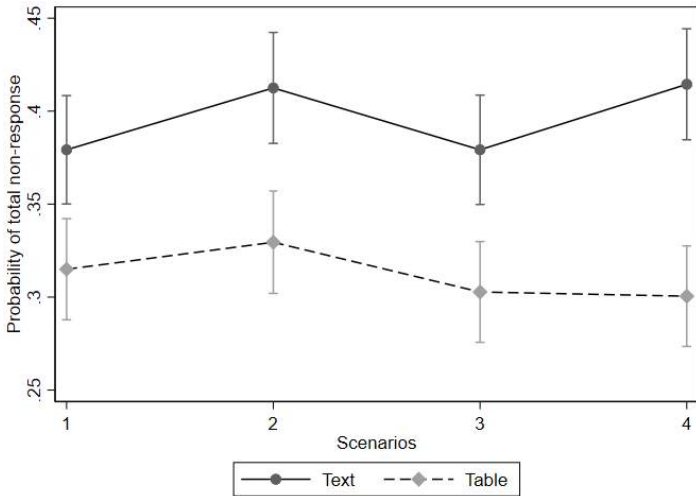
The total loss of information was, as expected, lower for the table format than for the text format ($b=-0.38$, $SE=0.08$, $p<0.001$). Again, the effect of presentation format on total non-response in the paired conjoint experiment was found to depend on the number of scenarios that the respondent had answered. Across all scenarios, total non-response was greater for the text format, but as the respondent evaluated more and more scenarios, the effect got stronger. Albeit the increased number of total non-response caused by the text format was not yet significant in the second and third scenario ($b_{\text{table} * \text{scenario } 2} = -0.07$, $SE=0.07$, $p=0.31$; $b_{\text{table} * \text{scenario } 3} = -0.06$, $SE=0.08$, $p=0.45$), by the fourth scenario this moderating effect became great enough to reach statistical significance ($b_{\text{table} * \text{scenario } 4} = -0.22$, $SE=0.08$, $p<0.001$), see Figure 10.

As expected, there was a negative effect of age on total non-response, indicating that older respondents found the conjoint experiments more demanding and adopted satisficing strategies more often compared to younger respondents (see Figure 8).



Notes. N=6,505 (scenario level).

Figure 9 Predicted probabilities of scenario on non-response over presentation format, for the paired conjoint experiment.



Notes. N=8,489 (scenario level).

Figure 10 Predicted probabilities effect of scenario on total non-response over presentation format, for the paired conjoint experiment.

Moderating effects. Potential moderating effects of education and age were also tested in the paired setting, with total non-response as the dependent variable, age and education as moderators, and gender as a control variable (see SOM S2.4., Figures S5 and S6, for an illustration of the moderating effects).

Education. In line with the results from the single conjoint experiment, the interaction term between education and presentation format was not significant in the paired conjoint experiment. Hypothesis 2c was, therefore, not supported.

Age. The results for age as a moderator were similar to those for education. The effect that presentation format had on total non-response did not on average significantly differ between age groups. There was a small and significant effect for 60-to-69-year old's of seeing the text format on total non-response (see Figure S6), which then disappeared for respondents older than 69 years. Overall, our results did not support the hypothesis that the effect of presentation format was moderated by age (H2d).

Response inconsistency and partial dimension reduction. In contrast to the single conjoint experiment, respondents in the paired conjoint experiment who were given the vignette in a table format yielded a statistically significantly stronger prediction for one of the dimensions in the first scenario (i.e., they had less response inconsistency) ($b_{\text{education: more than high school, not university}} = -0.14, SE=0.06, \chi^2(1, 1,932) = 5.13, p<0.05$).⁵ Furthermore, in the first scenario, the table format produced less measurement error ($\epsilon = 0.21, SE=0.00$) than the text format ($\epsilon = 0.23, SE=0.00, \chi^2(1, 1,932) = 6.80, p<0.01$) (see Table 7, column 1). In the subsequent scenarios, the parameter differences between the two presentation formats stabilized, with none of the parameters differing in scenario 2 (see Table 7, column 2), two differing in favor of the text format and one in favor of the table format in scenario 3 (see Table 7, column 3), and one favored the text and one favored the table format in scenario 4 (see Table 7, column 4).

A similar relationship was found for the error variance, where scenario 2 showed no difference in error variance, a statistically significant difference in favor of the table format in scenario 3, and no difference in scenario 4 (see Table 7, columns 2–4). Hence, the text format seems to have produced less response inconsistency in the first scenario; however, across all four scenarios, the response inconsistency analysis did not favor either of the presentation formats.

Furthermore, respondents did not evaluate the dimensions with increasingly less care over scenarios (i.e., adopted a partial dimension reduction behavior). This applied to both those who saw the questions in text format ($\epsilon_{\text{scenario 2} - \text{scenario 1}} = 0.01, \chi^2 = 0.63, p=0.43$; $\epsilon_{\text{scenario 3} - \text{scenario 2}} = 0.00, \chi^2 = 0.07, p=0.78$; $\epsilon_{\text{scenario 4} - \text{scenario 3}} = -0.01, \chi^2 = 1.31, p=0.25$) and those who saw them in table format ($\epsilon_{\text{scenario 2} - \text{scenario 1}} = 0.01, \chi^2 = 0.63, p=0.43$; $\epsilon_{\text{scenario 3} - \text{scenario 2}} = 0.00, \chi^2 = 0.07, p=0.78$; $\epsilon_{\text{scenario 4} - \text{scenario 3}} = -0.01, \chi^2 = 1.31, p=0.25$)

5 In the paired conjoint experiment, negative relationships between dimensions and the dependent variable were expected. Hence, the presentation format that produced the most negative coefficient was interpreted as the better performing format.

$\epsilon_{\text{scenario 1}} = 0.01, \chi^2 = 0.13, p=0.71; \epsilon_{\text{scenario 3} - \text{scenario 2}} = 0.02, \chi^2 = 2.37, p=0.12; \epsilon_{\text{scenario 4} - \text{scenario 3}} = 0.01, \chi^2 = 0.02, p=0.88$). Therefore, in contrast to the single conjoint experiment, neither of the presentation formats produced strong evidence of dimension reduction in the paired conjoint experiment.

However, overall, H2b was supported by the majority of the evaluation criteria on data quality. The table presentation format resulted in fewer refusals, break-offs, non-responses, and total non-responses, albeit no strong evidence for a stronger partial dimension reduction or inconsistency of responses was found for either of the two formats.

Table 7 Parameter differences between text and table format when predicting party choice with the attribute dimensions, for the paired conjoint experiment.

	Parameter differences (table – text)			
	Scenario 1	Scenario 2	Scenario 3	Scenario 4
Party level				
Status quo, tax and welfare	0.04 (0.05)	0.05 (0.06)	0.12* (0.06)	0.05 (0.06)
Less tax, less welfare	-0.02 (0.05)	0.02 (0.05)	-0.00 (0.05)	-0.03 (0.05)
Status quo of refugees	-0.03 (0.05)	0.04 (0.06)	-0.01 (0.06)	-0.01 (0.05)
More refugees	-0.10+ (0.05)	-0.01 (0.05)	-0.12* (0.05)	-0.12* (0.05)
Status quo of gender roles	-0.03 (0.05)	0.07 (0.06)	0.08 (0.06)	0.09 (0.05)
Traditional gender roles	-0.07 (0.05)	0.04 (0.05)	0.01 (0.05)	0.05 (0.05)
Party leader				
Hired unreported workers	-0.01 (0.05)	0.00 (0.05)	0.05 (0.05)	0.06 (0.05)
Drunk driving	0.01 (0.06)	-0.01 (0.05)	-0.07 (0.05)	-0.04 (0.05)
Female party leader	-0.06 (0.04)	-0.02 (0.04)	0.05 (0.04)	0.06 (0.04)
Education: Less than high school	-0.09 (0.06)	0.08 (0.06)	-0.03 (0.06)	0.07 (0.06)
Education: High school	-0.01 (0.06)	0.06 (0.06)	0.12* (0.06)	0.11+ (0.06)
Education: More than high school, not university	-0.14* (0.06)	0.04 (0.06)	0.03 (0.06)	0.03 (0.06)

	Parameter differences (table – text)			
	Scenario 1	Scenario 2	Scenario 3	Scenario 4
Constant	0.76*** (0.05)	0.95*** (0.06)	0.84*** (0.06)	0.91*** (0.05)
Error variance, table format	0.21*** (0.00)	0.23*** (0.00)	0.22*** (0.00)	0.22*** (0.00)
Error variance, text format	0.23*** (0.00)	0.22*** (0.01)	0.23*** (0.00)	0.23*** (0.00)
Error variance difference (table – text)	-0.02**	0.01	-0.01*	-0.01
χ^2 of difference	6.80	0.26	4.04	1.84
Observations	1,932	1,932	1,932	1,932

Notes. Regression coefficients from four ordinary least squares (OLS) regression equations, with clustered robust standard errors (SEs) in parentheses nested within respondents. Negative parameters indicate where the table format outperformed the text format in predicting the party choice, whereas positive parameters indicate where text outperformed the table format. Omitted dimensions were “more gender equality,” “more welfare, higher taxes,” “strive towards more gender equality,” “drunk driving scandal,” “male party leader,” and “education: university.” Only the participants who answered all four scenarios were included. See SOM S2.4., Table S15, for the parameters, separately for presentation format and scenario.

+p<0.1, *p<0.05, **p<0.01, ***p<0.001.

Summary of Results and Comparison to Shamon et al. (2019)’s Findings

In this paper, we presented a direct replication of the single conjoint presentation format experiment reported in Shamon et al. (2019), albeit with a few changes to the procedure, sample size, and analysis strategies. For a full description of differences compared to Shamon et al. (2019), see SOM S3.1–S3.2.

Our procedure makes some direct comparisons to Shamon et al. (2019) difficult. For example, our experiment evaluated only two presentation formats (text and table format), whereas Shamon et al. (2019) included an additional text format where the dimensions were underlined. We opted not to include the underlined presentation format because respondents have been reported to interpret the underlining of text in inconsistent ways (Reber, Schwarz & Winkielman, 2004), and underlining can result in more random measurement error (see Reber, Schwarz & Winkielman, 2004; Song & Schwarz, 2008).

Furthermore, presenting respondents with only four scenarios to evaluate (instead of 16, as in Shamon et al., 2019) made evaluations of consequent dimension

reduction unfeasible. However, decreasing the number of scenarios enabled us to heavily increase the statistical power of most of our other analyses (2,068 participants in the single conjoint experiment in our study compared to 498 in Shamon et al., 2019). Increased statistical power enabled us to identify whether Shamon et al.'s (2019) directional, albeit not statistically significant, effects of the text presentation format on decreased data quality were due to statistical power.

Despite some differences in design and analysis, the results in the single conjoint setting replicated several of the findings in Shamon et al. (2019). For example, despite our larger sample size, we also found no significant differences in terms of refusals or break-offs between the text and the table presentation format. Furthermore, Shamon et al. (2019) found no interaction effects between age and presentation format, which agrees with our findings.

However, some results found in this study did not replicate the findings reported by Shamon et al. (2019). For instance, in stark contrast to Shamon et al. (2019) who found no difference in partial dimension reduction and response inconsistency, we found that the table format statistically significantly outperformed the text format in less partial dimension reduction and less response inconsistency as the number of evaluated scenarios increased. Furthermore, Shamon et al. (2019) found no differences in the cost of administration, in terms of administration time, whereas our results favored the table presentation format. These differences may be due to the extra statistical power afforded by our sampling strategy.

In addition, Shamon et al. (2019) found that the text presentation format significantly showed decreased non-responses while the table format yielded fewer total non-responses, whereas we found no such differences.

The present paper, moreover, extended Shamon et al.'s (2019) work by conceptually replicating their experimental design in a paired conjoint experimental setting. Our conceptual replication produced even clearer evidence in favor of the table format. The table format outperformed the text presentation format by reducing the cost of administration and lowering refusal, break-off, non-response, and total non-response rates. The effect of presentation format on dimension reduction was, however, inconclusive.

Furthermore, the present paper extends Shamon et al.'s (2019) analyses by including two additional measurements of respondent experience, namely, respondents' evaluation of the questionnaire and participation in the waves of the Swedish Citizen Panel following the presentation format experiment. In contrast with other findings presented here, these additional measurements favored the text presentation format in terms of participation in subsequent waves but the table presentation format in terms of overall questionnaire evaluation in the single conjoint experiment. The presentation format had no significant effect on the evaluation of the questionnaire or participation in the panel waves following the experiment in the paired conjoint experiment.

Conclusion and Discussion

This paper investigated the impact that the presentation format (text or table) had on respondents' answering behavior by replicating Shamon et al.'s (2019) study on single conjoint experiments, as well as extending their work to also include paired conjoint experiments, where respondents state their preferences over two dimension sets/profiles.

Overall, the results in the present study favored the table over the text presentation format. As evidence of this, the table presentation format in both the single conjoint and the paired conjoint setting was found to statistically significantly outperform the text presentation format with regard to the cost of administration (i.e., the time it took respondents to evaluate the scenarios). However, a shorter administration time may, in fact, not be favorable if it is shorter because respondents answer faster by employing a suboptimal response process and satisficing strategies. Even though the respondents who were presented with the table format took less time to evaluate the scenarios, this shorter processing time did not clearly stem from less cognitive effort invested in the response.

Although respondents in the single conjoint setting produced stronger loadings on the dimensions in the first scenarios when reading the text instead of the table format, the respondents who read the text format suffered ever stronger partial dimension reduction (i.e., a decreasing impact of the dimensions and increased measurement error over the number of scenarios) as they evaluated more scenarios. In fact, by the fourth scenario, the table format had started producing stronger dimension loadings and significantly less measurement error than the text format. Hence, when respondents will evaluate only one scenario, the text format may be preferable, but as the number of scenarios increases, the table format seems to produce better, and more consistent, data quality. Our finding may have stemmed from the fact that respondents became fatigued more quickly by the text than by the table format, although the present study does not have the type of data that provide evidence for such a claim.

Similarly, respondents who saw the table presentation format in the paired conjoint setting evaluated the scenarios faster than respondents who read the questions in text format and did so without introducing partial dimension reduction or response inconsistency. Furthermore, in the paired conjoint setting, the table format outperformed the text format in other data quality measures, such as the number of refusals, break-offs, non-response, and total non-response. Overall, we found more distinct support for the table format in the paired conjoint setting compared to the single conjoint setting. The stronger evidence in the paired setting may be due to the presentation format having a greater impact when respondents evaluated two profiles or from the difference in topics between the single and the paired conjoint

experiments. Future studies that alternate topics on the single and paired conjoint settings to bring clarity on how sensitive the results are to the topics chosen.

The proposed theoretical benefit of the text format was that nesting the information within stories was thought to enhance respondents' understanding and empathy of the hypothetical situation. The increased understanding was, in turn, thought to increase the respondents' attention to the dimensions and increase the quality of the data. Furthermore, theoretically, respondents may be more likely to be accustomed to absorbing information in text paragraphs rather than tables. In contrast to these theories, our findings offer no support for any of these claims. Rather, respondents seem to connect to the information in the table emphatically and interpret the tables accurately, even when those tables are presented with two sets of profiles, which should have increased the complexity of the information to absorb.

Moreover, the present manuscript did not use any visual emphasis on the dimensions in the text vignettes (e.g., underlining, italicizing, or using bold fonts). Emphasizing the text that represented the dimension might have helped respondents to focus on the most relevant pieces of the vignette texts and could have made the text format perform better than what we found. However, we believe it to be unlikely that adding a visual emphasis would have negated our results because previous research has found that emphasis can make texts more difficult to read and understand (Reber, Schwarz, and Winkielman, 2004; Song and Schwarz, 2008), and emphasis can increase the time it takes respondents to evaluate conjoint vignettes (Shamon et al., 2019).

Counterintuitively, in the single conjoint setting, respondents were found to be more satisfied when receiving the table format, while not producing better data quality compared to the respondents who received the text format. By contrast, in the paired conjoint setting, respondents who were given the table format were not more satisfied with the questionnaire but produced statistically significantly better data quality compared to the respondents who were given the text format. A potential explanation for this dissimilarity may be that the most dissatisfied respondents stop filling out the questionnaire before getting to the questionnaire evaluation questions, leading to artificially greater satisfaction ratings for the worst performing presentation format (greater, because only more satisfied respondents answer the questionnaire evaluation questions). However, whereas we did find greater satisfaction among table format respondents in the single conjoint setting, we did not observe more refusals, break-offs, non-response, or total non-response for either of the presentation formats in the single conjoint.

We did observe more refusals, break-offs, non-response, and total non-response, but no differences in respondent satisfaction, in the paired conjoint setting. The only instance where break-offs, refusals, non-responses, and total non-response could artificially produce the satisfaction ratings we found would be if

respondent satisfaction among those presented with text format started at lower levels than among those receiving the table format. The artificial increase in satisfaction afforded by the break-offs, refusals, non-responses, and total non-response would then bring the mean satisfaction with text format to the same levels as satisfaction with the table format. However, random assignment of the two formats should limit such an outcome. Perhaps, rather than thinking of the findings as counterintuitive, the results of this study indicate that respondent satisfaction and data quality are two distinct phenomena, each offering different insights and advice for survey researchers. Survey researchers should be interested in both phenomena, but if forced to choose, better data quality should be preferred over respondent satisfaction, especially as respondents seem able to be unsatisfied with a questionnaire while still more likely to complete each conjoint evaluation.

Lastly, both our and Shamon et al.'s results (2019) were based on online convenience samples. Online convenience samples have been found to be more suitable for generalization of treatment effects than, for example, student samples (e.g., due to being more diverse in educational attainment, age, gender, and income of the respondents, see Berinsky, Huber & Lenz, 2012). Hence, the non-difference between text and table formats found in Sauer et al., (2020) could be due to their student sample being more accustomed to reading lengthier text paragraphs than a general population sample. In contrast, the chance remains that our self-selected panelists and those in Shamon et al. (2019) may be more literate in reading tables than the general population, potentially producing the outperforming of the table format in our experiments. Future research should attempt to replicate similar presentation format experiments among probability sampled respondents.

In the meantime, based on the results of this study and those reported in Shamon et al. (2019), we conclude that respondents simply seem less likely to resort to satisficing strategies when evaluating conjoint experiments using a table presentation format than when evaluating them in a text format. For now, we argue that a table presentation format is to be preferred when designing conjoint experiments distributed online.

References

- Auspurg, K., & Hinz, T. (2015). *Factorial Survey Experiments (Quantitative Applications in the Social Sciences)*. Thousand Oaks, CA: SAGE Publications.
- Auspurg, K., & Jäckle, A. (2017). First equals most important? Order effects in vignette-based Measurement. *Sociological Methods & Research*, 46(3):490–539.
- Bansak, K., Hainmueller, J., Hopkins, D., & Yamamoto, T. (2021). Beyond the breaking point? Survey satisficing in conjoint experiments. *Political Science Research and Methods*, 9(1), 53–71. Doi:10.1017/psrm.2019.13

- Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis*, 20(3), 351–368. <https://doi.org/10.1093/pan/mpr057>
- Hainmueller, J., Hangartner, D., & Yamamoto, T. (2015). Validating vignette and conjoint survey experiments against real-world behavior. *Proceedings of the National Academy of Sciences*, 112(8), 2395–2400.
- Hainmueller, J., Hopkins, D. J., & Yamamoto, T. (2014). Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Political Analysis*, 22(1), 1–30.
- Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology*, 5(3), 213–236.
- Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, 50(1), 537–567.
- Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly*, 51(2), 201–219.
- MacDonald, K. (2016). Group comparisons in structural equation models: Testing measurement invariance. *The Stata Blog*. Accessed December 17th, 2020, at <https://blog.stata.com/2016/08/23/group-comparisons-in-structural-equation-models-testing-measurement-invariance/>
- Mahon-Haft, T. A., & Dillman, D. A. (2010). Does visual appeal matter? Effects of web survey aesthetics on survey quality. *Survey Research Methods*, 4(1), 43–59.
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251).
- Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? *Personality and Social Psychology Review*, 8(4), 364–382.
- Roberts, C., Gilbert, E., Allum, N., & Eisner, L. (2019). Research synthesis: Satisficing in surveys: A systematic review of the literature. *Public Opinion Quarterly*, 83(3), 598–626.
- Sato, H., Kubo, M., & Namatame, A. (2007). Video-based conjoint analysis and agent based simulation for estimating customer's behavior. *International Conference on Intelligent Data Engineering and Automated Learning* (pp. 1102–1111). Springer, Berlin, Heidelberg.
- Sauer, C., Auspurg, K., & Hinz, T. (2020). Designing Multi-Factorial Survey Experiments: Effects of Presentation Style (Text or Table), Answering Scales, and Vignette Order. *Methods, data*, 14(2) Pages. <https://doi.org/10.12758/MDA.2020.06>
- Schuman, H., & Presser, S. (1996). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Sage.
- Shamon, H., Dülmer, H., & Giza, A. (2019). The Factorial Survey: The Impact of the Presentation Format of Vignettes on Answer Behavior and Processing Time. *Sociological Methods & Research*. <https://doi.org/10.1177/0049124119852382>
- Simon, H. A. (1957). *Models of man; social and rational*. Wiley
- Song, H., & Schwarz, N. (2008). If it's hard to read, it's hard to do: Processing fluency affects effort prediction and motivation. *Psychological Science*, 19(10), 986–988. <https://doi.org/10.1111/j.1467-9280.2008.02189.x>
- Tukey, J.W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

Improving Anchoring Vignette Methodology in Health Surveys with Image Vignettes

*Mengyao Hu¹, Sunghee Lee¹, Hongwei Xu²,
Roberto Melipillán³, Jacqui Smith¹ & Arie Kapteyn⁴*

¹ *University of Michigan-Ann Arbor*

² *Queens College, New York*

³ *Universidad Del Desarrollo, Chile*

⁴ *University of Southern California*

Abstract

The anchoring vignette method is designed to improve comparisons across population groups and adjust for differential item functioning (DIF). Vignette questions are brief descriptions of hypothetical persons for respondents to rate. Although this method has been adopted widely in health surveys, there remain challenges. In particular, vignettes are complex, increasing survey time and respondent burden. Further, the assumptions underlying this method are often violated. To overcome such challenges, this paper introduces an innovative technique, namely image anchoring vignettes, conveying vignette information with varying health levels in images. We conducted a cross-cultural experimental study to examine the performance of image and standard text vignettes in terms of response time, how well they satisfy the assumptions, and their DIF-adjusting quality using a confirmatory factor analysis. The study revealed that respondents can better differentiate the intensity levels of the three vignettes in the image vignette condition, compared to text vignettes. Response consistency assumption appears to be better satisfied for image vignettes than text vignettes. Using well-designed image vignettes greatly reduces survey time without losing the DIF-adjustment quality, indicating the potential of image vignettes to improve overall efficiencies of the anchoring vignette method. Improving vignette equivalence (i.e., minimizing different interpretations of vignettes by different groups), remains a challenge for both text and image vignettes. This study generates new insights into the design and use of image anchoring vignettes.

Keywords: Differential item functioning; Anchoring vignettes; Image vignettes; Cross-cultural comparisons; Self-assessments of health



Self-assessed questions on health are good predictors for mortality and morbidity (Idler & Benyamini, 1997; DeSalvo et al., 2005). Self-assessment health questions often use Likert-type rating scales to measure respondents' attitudes, knowledge, perceptions, and behavior (Krosnick & Abelson, 1992; Lee, Jones, Mineyama, & Zhang, 2002). Ideally, responses obtained from these questions reflect only respondents' true state. This, however, is not always the case. In fact, answers to self-assessment questions reflect both respondents' true state and how they use the scales, a phenomenon known as response-category differential item functioning (DIF) (King, Murray, Salomon, & Tandon, 2004; King & Wand, 2007). As described in King and Wand (2007), DIF refers to situations when respondents from different backgrounds map the same state onto the scales in different ways.

Figure 1 (adapted from Hu, Lee, & Xu, 2018) illustrates a cross-cultural study example of DIF to a self-assessed pain question on an ordinal response scale from "None" to "Extreme". In this example, cultural groups, A and B, use different cut points for a given response category. Assume that two respondents, one from A and one from B, have the same true pain level, both falling on the vertical dashed line. Despite their identical pain levels, the respondent from A will select "Mild," and the respondent from B will choose "Moderate". If this DIF is not accounted for, simple between-culture comparisons will erroneously conclude that the Culture B respondent experiences a higher level of pain (Hu et al., 2018).

An adjustment method for such DIF issues is to use anchoring vignettes (AV), which have been used in multiple national and international health surveys including the Health and Retirement Study (HRS) and the Survey of Health, Ageing and Retirement in Europe (SHARE). The AV approach typically involves two components: a self-assessment question and (typically multiple) anchoring vignette questions. First, respondents are asked to report their own status. For example, in a health survey, a typical self-assessed pain question is: Overall, in the last 30 days, how much pain or bodily aches did you have? The second component consists of vignette questions, each in a few sentences describing a hypothetical person's situation related to the construct measured, and respondents are asked to rate the vignette person. For example, a vignette used in HRS asks, "*Paul has a headache once a month that is relieved after taking a pill. During the headache he can carry on with his day-to-day affairs. Overall, in the last 30 days, how much of a problem did Paul have with bodily aches or pains?*". Usually, more than one vignette question describing varying intensity levels of the measured construct (e.g., low, moderate, and high levels) are asked (see Appendix 1). The vignette ratings can serve

Direct correspondence to

Mengyao Hu, Survey Research Center, Institute for Social Research,
University of Michigan, 426 Thompson St., Ann Arbor, MI, USA
E-mail: maggiehu@umich.edu

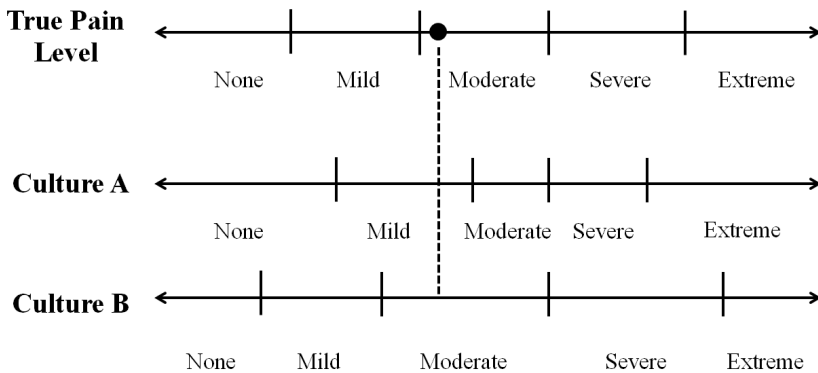


Figure 1 DIF for cross-cultural studies. Adapted from Hu, Lee & Xu (2018). The horizontal lines with arrows indicate the continuous scales of the domain (pain level). The short vertical lines indicate the cut points respondents used to answer the self-assessment question. The vertical dashed line indicates respondents responses to self-assessment questions. If a respondent's pain level falls on that line, it indicates that they have the same true pain level.

as benchmarks for the actual unobserved self-assessed pain level that researchers intend to measure.

The successful use of anchoring vignettes depends on two key assumptions: response consistency (RC) and vignette equivalence (VE). RC requires respondents to rate vignette persons in the same way as they would rate themselves (King et al., 2004). VE assumes that vignette descriptions are perceived similarly across respondents (King et al., 2004), essentially requiring vignettes to provide the same stimuli across respondents.

Promises and Pitfalls of the Current Anchoring Vignette Approach

Anchoring vignettes (AV) have been reported in many studies as a promising tool to correct for DIF (e.g., Mojtabai, 2015; Murray et al., 2002). Despite its promise, studies of the effectiveness of the standard AV (which rely on verbal descriptions of the vignette persons) have yielded mixed results. While some studies have found that text vignettes can effectively correct for DIF (Dowd & Todd, 2011; Van Soest, Delaney, Harmon, Kapteyn, & Smith, 2011), other studies have reported that text vignettes do not necessarily provide comparable results among population groups (e.g., Grol-Prokopczyk et al., 2015). Previous studies have also shown that RC and

VE assumptions can be violated in different domains (Bolt, Lu, & Kim, 2014; Ferrer-i-Carbonell, Van Praag, & Theodossiou, 2011; Kapteyn, Smith, Van Soest, & Vonková, 2011; Rice, Robone, & Smith, 2012).

The assumption violations are likely due to several practical challenges related to the AV design [see also Hu and colleagues (2018)]. The first and most obvious challenge concerns question difficulty (Hopkins & King, 2010). Unlike typical survey questions that ask respondents to rate their own status, AV require respondents to imagine hypothetical persons based on verbal descriptions and to shift their focus from themselves to rate the status of these imagined hypothetical persons, placing greater cognitive burden on respondents. The second challenge is a substantial increase in survey time. Given that vignettes are designed to describe hypothetical situations, one single vignette often contains much more text than other typical survey questions (Hu and Lee, 2016). In addition, because usually more than one vignette is used per domain (e.g., pain), the use of AV may require a non-trivial amount of response time (Hirve et al., 2013; Hopkins & King, 2010; King et al., 2004). Third, the use of AV in cross-cultural research raises yet another issue with text vignettes: measurement inequivalence, where respondents with different cultural background may understand vignette descriptions in systematically different ways. One source that can lead to measurement inequivalence is questionnaire translation. Poor translation can directly influence respondents' interpretation of the vignettes, leading to violation of the VE assumption. Another critical challenge is the specific content to include in vignette descriptions. As acknowledged by Kapteyn et al. (2011), it is difficult to write vignette descriptions that are as "comprehensive" as what respondents know about their own state (Kapteyn et al., 2011). This indicates that respondents may rate themselves using criteria different from those they use for vignettes, resulting in violation of the RC assumption. VE can also be violated if respondents interpret the vignette descriptions in different ways. The potential for this problem is even greater in cross-cultural research where the challenges of designing equivalent and comparable vignettes are increased.

Although previous literature has greatly emphasized the importance of the design and pretesting of text AV, no clear design guidelines have been established to address the above limitations and practical challenges.

Image Anchoring Vignettes

As a potential remedy to the limitations of text AV, we propose in this study to use visual AV with well-designed and carefully-selected images, i.e., image vignettes. With the technical development of internet, image vignettes have gained increasing popularity in survey research, especially in studying attitudes and sensitive questions (Naylor et al., 2014; Groot et al., 2020). To the best of our knowledge, this study is the first research that incorporates visual methodology with AV techniques.

Mechanisms of information processing of visual vs. verbal stimuli have been discussed in previous studies but there are no consensus conclusions. Some studies report similar processing of visual and verbal information in “a functional unitary system that is directly accessed by both visual objects and words” (Caramazza, 1996). In contrast, some other studies have shown that visual and verbal information are processed differently and “creating separate semantic representations” (Glaser, 1992; Glaser & Glaser, 1989; Schlochtermeier et al., 2013). For example, information processing of images is reported to be connected to activation of the right brain hemisphere (Grady et al., 1998; Naspetti et al., 2016), and activation of the left hemisphere is found to be associated with text information processing (Sevostianov et al., 2002). Despite the inconclusive results of the mechanisms of information processing, a common finding reported in previous studies is the “processing superiority” of images as compared to text information (Azizian et al., 2006, Schlochtermeier et al., 2013). As reported in Schlochtermeier et al. (2013), images lead to faster and a more direct access to meaning. In comparison, texts require “additional translational activity at the representational level” to access the semantic system (Schlochtermeier et al., 2013).

Given the reported processing superiority of image processing, the image AV strategy may lead to several potential advantages. First, images may require less cognitive effort to process than do text descriptions. Compared to texts, images are processed in a quicker and more automatic way, allowing respondents to form more “direct” connections between images and their meaning (Luna & Peracchio, 2003; Paivio, 2013; Townsend & Kahn, 2014). In the case of AV (which require imagining hypothetical persons), the use of images is advantageous for both low-literacy respondents and those who are unable to create mental images based on text vignettes. For these respondents, the saying “A picture is worth a thousand words” is particularly relevant considering the challenge of reading through the lengthy text descriptions to understand the vignette scenario (Hibbing & Rankin-Erickson, 2003).

In addition to ease of understanding, because respondents can process information shown in image vignettes relatively quickly, we expect that the use of image vignettes will reduce respondents’ cognitive burden and overall survey time. In turn, these two aspects could contribute to improving survey data quality by reducing survey break-offs and respondents’ satisficing behavior.

A second potential advantage of image vignettes is that they might help satisfy the measurement assumptions. For example, it has been found that first names used in text vignettes (e.g., “Alice falls asleep easily at night...”) can lead to respondents’ inferences about that person’s characteristics, such as age, gender and racial/ethnic information (e.g., Jürges & Winter, 2013). If respondents from different groups perceive the vignette person as having different characteristics, VE is likely to be violated. This may be of less concern in well-designed image vignettes where

the physical characteristics of the vignette person are clearly presented, limiting the possibility of different interpretations. Note that the performances of image vignettes can largely depend on how they are designed. Some design features may be associated with different interpretations of the vignette person, e.g., respondents with different age and gender may view a vignette person with tattoos, piercings, and unnaturally colored hair differently. While it is true that not all image vignettes will help satisfy the measurement assumptions, in this study, we aim to investigate: with carefully designed image vignettes on health domains, whether image vignettes could help with measurement assumptions, compared to text vignettes.

Because there are no prior studies on the use of image anchoring vignettes, it remains an open question whether this approach can remedy limitations of current text vignettes. To fill this gap, this paper aims to evaluate the use of image AV as an alternative to text vignettes and to compare the performance of image and standard text vignettes in terms of response time, how well they satisfy the RC and VE assumptions, and their ability to reduce measurement errors in a confirmatory factor analysis (CFA) framework. In this paper, we focused on four health domains – sleep, affect, mobility, and pain – which are known to be subject to DIF (e.g., d’Uva, O’Donnell, & Van Doorslaer, 2008). We have three research questions (RQ).

RQ1: Will image AV reduce response time, compared to text AV? This research question will be addressed by analyzing survey time associated with text and image AV using time stamp data.

RQ2: Will image AV better meet AV measurement assumptions compared to text AV? This research question will be addressed by examining both VE and RC assumptions for text and image AV.

RQ3: In a confirmatory factor analysis (CFA) framework, a.) we will investigate whether a model of latent health based on image or text AV-adjusted scores will show better fit compared to a model based on unadjusted self-reported scores, and b.) whether a model based on image AV-adjusted scores will have similar or better fit compared to a model based on text AV-adjusted scores, i.e., will image AV adjustment achieve similar or better measurement error-reduction, compared to text AV?

Methods

Design of Image Vignettes

Prior to designing the image vignettes, we established criteria for image selection or creation. A three-step approach was used to develop these criteria: specifically, we 1) thoroughly examined critical elements of the four health domains, 2) identified common elements applicable across groups (e.g., arm pain) based on the litera-

ture review, and 3) based on the elements identified, we selected or designed images with these elements at different intensity levels for each domain (e.g., from no pain to extreme pain). Based on the developed criteria, images were then selected from commercial websites of images and photos (e.g., www.istockphoto.com/). In situations where, for a given health domain, no images meeting the criteria were found on those websites, we 1) recruited volunteers from different platforms (e.g., friends or family members) to serve as models in the photos, 2) obtained each volunteer's consent to take a photo and to use it in this study, and 3) took the photo and edited them. To remove potential confounding effects of various image elements, such as background, size, resolution, and color balance, the selected images or photos were further edited by students with expertise in image-editing.

The ultimate goal of the image vignette design for the current study was to have three well-designed image vignettes per domain. For the purpose of selecting the most comparable images across cultures, we first designed six images for each characteristic: two images for each intensity level (e.g., two no/low pain, two moderate pain and two extreme pain vignettes) per design condition, and eventually selected three out of the six for each condition in the pretest. The selected images (see Appendix 1) were then used in the web survey experiment as described below¹.

Pretesting

The pretest was conducted through Amazon Mechanical Turk (MTurk), where we posted the survey announcement, also known as Amazon's human intelligence tasks (HITs). Eligible respondents can browse the HITs and decide if they would like to take the survey or not. The announcement contains a link to the pretest survey, which was programmed with Qualtrics. The pretest was open to U.S. workers who were 18 or older. A \$0.45 incentive was offered for each completed survey. To recruit respondents of all age groups, toward the end of the data collection, we posted a HIT open only to older respondents with the same incentive. In total, 201 respondents completed the pretest survey, about half of them aged 50 years or older. The main criteria applied to evaluate and select proper images was based on whether respondents could correctly rank order vignettes as expected. This method was first used by World Health Organization (WHO) in their pretesting of anchoring vignettes (Murray et al., 2003). For the two sets of image options, the image with the higher correct ranking rate (the percentage of respondents who correctly

1 In designing image vignettes, two different conditions (e.g., male and female) were designed for each domain. Respondents assigned to the image vignette conditions were randomly assigned to the two design conditions. This paper focuses only on the comparison between text and image vignettes, and evaluations on how image vignette design features influence anchoring vignette methodology are discussed elsewhere.

ranked the vignette series) was selected. The final correct ranking rates ranged from about 80% to 97% across all health domains.

Web Survey Procedure

The main data collection was based on a web survey using a non-probability online panel. Respondents from four different racial/ethnic groups – Non-Hispanic (NH) white, NH black, English-speaking Hispanic and Spanish-speaking Hispanic – were recruited through Qualtrics' online survey panel, which partners with over 20 Web-based panel providers to supply diverse, quality respondents (more information about Qualtrics survey panel, see also Holt & Loraas, 2019; Ibarra et al., 2018). The reason for including these groups is that race/ethnicity and language are proxies of cultures (Davis et al., 2019; Lee et al., 2014; Lee et al., 2017) and are known to influence respondents' self-reporting of their health status (McCarthy, Ruiz, Gale, Karam, & Moore, 2004; Lee et al., 2014). For example, Hispanics have been shown to conceptualize health differently than non-Hispanic Whites as they "include non-medical aspects, such as spiritual and social wellbeing, in addition to medical conditions that non-Hispanic Whites consider the most critical element for assessing health" (Lee et al., 2014). Language can also influence respondents' reporting of their health status, e.g., Lee and colleagues examined Hispanics' self-reported health by interview language and found that the difference was primarily due to Hispanics interviewed in Spanish (Lee et al., 2014). Respondents from each racial/ethnic group were randomized into three conditions: the standard text vignette condition and two image vignette conditions that differed in the vignette persons' characteristics (See Appendix 2 for a flowchart of the experimental conditions and assignments). Robustness of randomization was examined, and results show that there are no significant socio-demographic differences across the experimental conditions (Supplemental Table 5), suggesting that the randomization works well.

For the text vignette condition, we adapted the text vignette descriptions from those widely used in many major surveys (e.g., HRS). Each domain had a series of three vignettes, describing different intensity levels of the measured construct: low, moderate and high (e.g., from least to most pain). For the image condition, we used the image vignettes designed and selected in the pretest with three vignettes per condition, depicting three levels of difficulty/intensity of symptoms in each domain (see Appendix 1). The introduction to the vignette questions also followed the standard approach used in earlier surveys such as HRS. We randomized the order of the domains and of the three vignettes per domain presented to respondents in order to isolate question order effects. Besides self-assessment and vignette questions, the study also included responses to objective questions regarding these health

domains, time stamp data, and respondents' demographic and socio-economic information.

In translating the instrument into Spanish for Spanish-speaking Hispanics, this study followed the set of best practices developed by the United States Census Bureau (Pan & De La Puente, 2005) and the Cross-Cultural Survey Guidelines developed by the survey research center at the University of Michigan (Mohler et al., 2016). Translation was conducted by the translation team of HRS. The translated questionnaire was then reviewed and tested by 20 bilingual speakers who are native Spanish speakers and are also fluent in English.

The online survey questionnaire was programmed in Qualtrics. The Qualtrics online panel team sampled respondents from their panel. Except for Hispanics speaking Spanish, around 750 respondents were sampled for each of the three other race/ethnic groups. Each of the three sampled subgroups had nearly equal proportions of 1) male and female, 2) below or equal to high school education and higher than high school education, and 3) respondents aged 18-49 or 50 and over. For Spanish-speaking Hispanics², 889 respondents were sampled with about 43% male respondents. Detailed information of the sample profile is presented in Table 1. In conducting this experiment, we implicitly make the stable unit treatment value assumption (SUTVA) that the outcome for one respondent is unaffected by the assignment of treatments to the other units. This assumption is likely to have been met in our study given Qualtrics' large pool of respondents and our duplicate check on respondents' IP addresses.

Email invitations were sent to selected respondents, with the link to the survey included in the email. Respondents from each racial/ethnic group were randomly assigned to one of the three vignette type conditions, one text condition and two image conditions.

2 Due to the difficulties in recruiting Spanish-speaking Hispanics, Qualtrics collected more respondents for this group in order to meet the targeted number of male Spanish-speaking Hispanics who were 50 and above and had education equal to high school or below.

Table 1 Respondents' characteristics.

	White (n=760) %	Black (n=750) %	Hispanic- English (n=750) %	Hispanic- Spanish (n=889) %
Male	50.39	50.00	50.00	42.52
Age				
Age 18 – 29	14.34	22.80	22.13	21.37
Age 30 – 49	33.68	25.73	26.53	35.77
Age 50 – 64	30.13	36.27	34.53	33.52
Age 65 and above	21.84	15.20	16.80	9.34
More than high school	49.47	50.00	50.00	57.82
Married	53.42	36.67	50.93	54.78
Employed	50.92	52.00	56.13	57.14
Income				
Income below \$40,000	35.00	35.87	33.07	34.76
Income between \$40,000 - \$69,999	33.95	42.93	41.33	45.67
Income \$70,000 or more	31.05	21.20	25.60	19.57

Analysis Strategy

We first examined the distributions of the self-assessment and vignette questions by vignette type for each domain descriptively. We then examined whether and to what extent the self-assessments were affected by DIF following previous literature studying measurement errors in self-assessed health (Yan & Hu, 2018). Specifically, since self-assessments of health are correlated with objective health conditions (Idler & Kasl, 1995), we take advantage of this relation to gain insights on how DIF affects respondents' uses of the scales. We constructed a measure of objective health for each domain using respondents' own answers to a series of factual questions asking about health conditions for each domain. We then standardized the number of health issues (e.g., the number of mobility issues) within each racial / ethnic group. The resultant standardized score reflects the number of standard deviations above or below the racial/ethnic subgroup mean, where a value of 0 stands for the subgroup average. Negative values of health scores denote better health than the subgroup average (i.e., respondents reported fewer health conditions) whereas positive values indicate worse health than the racial/ethnic subgroup average (i.e., respondents reported more health conditions). For each category selected

on the self-assessment question, we computed the mean of the standardized scores and compared them across different racial / ethnic groups.

We then examined RQ1 to RQ3 as described below. Note that in examining RQ1 to RQ3, the variables were not standardized.

RQ 1. To evaluate whether image vignettes can reduce survey time compared to text vignettes, we analyzed the survey time using time stamp data. The mean response time was compared between the text and image vignette types. To formally test the effects of vignette types on survey time, for each domain, we fit multilevel linear regression models with random intercepts. The log-transformed response time was used as the outcome, given that time is right skewed. In this model, Level 1 corresponds to vignette questions, and Level 2 corresponds to respondents. Level 1 covariate was vignette type (image vs. text vignettes) and Level 2 covariates included respondents' demographic and socio-economic variables. Results of the multilevel model can be found in Appendix 3 (Supplemental Table 6). Given that it is hard to ascertain whether respondents were completing the online survey from beginning to the end in one sitting or took temporarily breaks – e.g., checking emails and browsing other web tabs, we employed a two-step procedure to identify response time outliers. First, based on the response time distribution, we used 15 minutes (i.e., 900 seconds)³ per vignette question as a threshold to identify those who might took a break during the survey completion. Second, we examined distributions of random effects and residuals of the multilevel models described above. Using histograms and Q-Q plots, outliers on these parameters were inspected visually. In total, the first step identified four response time outliers for pain domain, two outliers each for sleep and mobility domains and six outliers for affect domain were identified and excluded from this analysis. The second step did not identify any outliers.

RQ 2. We compared image and text vignettes in terms of how well they satisfy the two measurement assumptions – VE and RC. Below we describe approaches for each of the two assumption-testing.

RQ 2a (Test for VE). Two tests of VE were conducted. The first one is referred as correct rank ordering test, which examines whether respondents could correctly rank order vignettes based on their intensity level. Several previous studies refer to this test as a weak test for VE, stating that correct rank-ordering is a “necessary but not sufficient” condition for VE (e.g., Grol-Prokopczyk et al., 2015; Kristensen & Johansson, 2008), given that if VE is fulfilled through effective vignette design, respondents should agree on the ranking of the vignettes.

It is possible that respondents may rate two or three vignettes identically. For example, if a respondent has a very high threshold for what is “mild” pain, that respondent may rate the first two vignettes (low and moderate pain) or all

3 As a sensitivity analysis, we also performed the analysis with 5 minutes and 10 minutes thresholds to identify response time outliers, which gave consistent results.

vignettes as no pain. This is referred to as “ties” in vignette-ratings. Although it is possible that a respondent may have *true* ties for all three vignettes (i.e., view the three vignettes as having similar intensity levels and rate them identically), this is unlikely given the differences among the intensity levels in the vignette design. Thus, here we only consider two kinds of ties: 1) ties between the first two vignettes (low and moderate intensity) and 2) ties between the last two vignettes (moderate and high intensity).

The second test for VE was a statistical test conducted following Grol-Prokopczyk (2018). This method was first developed by d’Uva et al. (2011) and applied in many other studies (Grol-Prokopczyk, 2018; Grol-Prokopczyk et al., 2015; Molina, 2016). The rationale behind this test is that if respondents view each vignette in the same way (VE), the distance between any two vignettes on the latent dimension should be the same for all respondents (d’Uva, Lindeboom, O’Donnell, & van Doorslaer, 2011). The test is based on a likelihood-ratio (LR) test of two nested models. Both models are variations of the hierarchical ordered probit (HOPIT) model. Below we list the key differences between the two models. The first model, Model (A)⁴, predicts a respondent’s perceived location of vignettes:

$$V_{ij}^* = \alpha_j + \varepsilon_{ij} \quad (\text{A})$$

where V_{ij}^* is respondent i ’s perceived location of vignette j on the latent dimension, α_j is a constant term and ε_{ij} is the random error term that is assumed to be normally distributed with mean zero and variance one. For one of the vignettes in a domain (the reference vignette), α is set to 0 for model identification. The cut points (τ) for the vignettes are modeled in the same way as in the HOPIT model. Note that Model A does not include covariates to predict perceived vignette locations on the latent dimension. This is consistent with VE, namely that respondents’ perceptions of vignettes do not depend on their background and are constant across different population groups.

In the less restrictive Model B, a vector of covariates, \mathbf{X}_i , is added to Model A to predict the perceived vignette locations. In this study, \mathbf{X}_i includes marital status, employment status, age, gender, education, income level, and racial/ethnic group.

$$V_{ij}^* = \alpha_j + \lambda_j \mathbf{X}_i + \varepsilon_{ij} \quad (\text{B})$$

Since this model is not identified, one needs a normalization. For one of the vignettes (the reference vignette), both α and λ_j are set to zero for identification. If VE is satisfied, λ_j will be 0 for each j . Model A is nested in Model B and if VE is satisfied, the LR test will not reject Model A. If, however, the LR test rejects Model

4 In describing the models, we used the same notation as Grol-Prokopczyk & Carr (2017).

A, it indicates that respondents with different characteristics perceive the severity of the vignettes differently. The estimated coefficient vector λ_j will indicate which covariates are driving the violation of VE.

RQ 2b (Test of RC). Our test of RC was conducted following Grol-Prokopczyk et al. (2015). This test was based on visual comparisons of two sets of predicted cut points. One set was generated from vignettes only, based on Model A as in the tests of VE. The other set was generated from self-assessments based on Model C below, which uses objective health measures to predict the self-assessments.

$$Y_i^* = \mu + \beta \mathbf{W}_i + \varepsilon_i \quad (\text{C})$$

where Y_i^* is respondent i 's true score on the latent dimension in the measured domain, μ is a constant term and ε_i is a random error term that is assumed to be normally distributed with mean zero and variance one. \mathbf{W}_i is a vector of covariates consisting of the objective measures. The cut points are modeled in the same way as in Model A. The predicted mean cut points from the two models were then graphed in a figure for visual comparisons. The RC test basically compares the shape (Grol-Prokopczyk et al., 2015) of the two sets of cut points. A similar shape would indicate that respondents had similar standards when rating vignettes and rating themselves (RC). As mentioned in Grol-Prokopczyk (2018), this test can be viewed only as suggestive. The objective measures used in this study include: whether respondents have seen a doctor about their difficulties with sleep, whether respondents on average sleep less than 7 hours or over 9 hours each day, a sleep quality score⁵, total pain index⁶, number of mobility activities that respondents have difficulty with, number of chronic health conditions, and the Kessler Psychological Distress Scale (K6) (Kessler et al., 2002).

RQ 3. The self-assessments for the health domains have often been used in a confirmatory factor analysis (CFA) framework to measure latent overall health. To examine whether AV-adjustment can reduce measurement errors in self-assessments, following Weiss & Roberts (2018), we compared the model fit of the CFA using original responses with the CFA using text / image AV-adjusted scores. If the use of AV-adjusted scores can correct DIF, we would expect the models with AV-adjusted scores to have better fit (RQ 3a; see also Weiss & Roberts, 2018). To evaluate whether image AV can achieve similar or better DIF-correction compared to text AV (RQ 3b), we also compared the magnitude of improvement compared to CFA with original self-reports, for both image and text AV-adjustment.

5 The sleep quality score was constructed based on responses to three sleep questions, asking respectively whether and how often respondents 1) have trouble falling asleep, 2) wake up several times at night, and 3) wake up earlier than planned at night and are unable to fall asleep again.

6 The total pain index was constructed following Ray et al. (2009).

The AV-adjusted scores were calculated using the non-parametric approach, following previous literature (Wand et al., 2011). In situations where respondents have ties in their AV-rating or inconsistent AV orders from researchers' expected order (i.e., order violations), the non-parametric method will result in an interval instead of a number for these respondents. Following the recommendations in previous literature (Kyllonen & Bertling, 2014; Primi, Zanon, Santos, De Fruyt, & John, 2016; Weiss & Roberts, 2018), the lower bounds of the intervals are chosen as the adjusted scores for respondents with ties or order violations. Model fit criteria including Comparative-Fit-Index (CFI), Tucker–Lewis index (TLI), and a Root Mean Square Error of Approximation (RMSEA) and 90% confidence interval (CI) of RMSEA are used to compare the models (Schreiber et al., 2006). A CFI greater than 0.95 and a TLI greater than 0.95 are considered as acceptable model fit (Hu & Bentler, 1999). A RMSEA less than or equal to 0.05 is considered as good fit, and less than or equal to 0.08 is considered as moderate fit (MacCallum, Browne & Sugawara, 1996). For the 90% CI of RMSEA, ideally the lower value should be less than 0.05 and the upper value less than 0.08 (MacCallum, Browne & Sugawara, 1996; Schreiber et al., 2006).

Results

Descriptive Analysis

We first examined the distributions of the self-assessment and vignette questions by vignette type for each domain. Figure 2 shows the distribution for the pain domain. Similar patterns were found for other domains. As expected for a properly randomized design, for each domain, the distributions for the self-assessment questions do not differ by vignette type-text or image vignettes. Comparing vignette distributions by vignette type, in general, the intensity levels of the image vignettes can be better differentiated than those of the text vignettes.

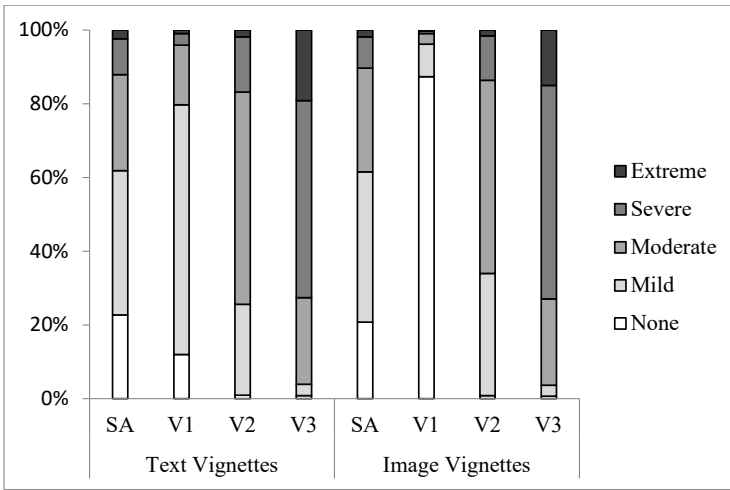


Figure 2 Responses to pain self-assessment (SA) and difficulty/intensity questions for three vignettes (V1 = none/mild; V2 = moderate; V3 = severe/extreme).

DIF Evaluation

We then examined whether DIF was present in the self-assessments⁷. Figure 3 displays the mean standardized number of mobility issues by reported response categories of self-assessed mobility. For all four racial / ethnic groups, the mean standardized scores are negative for those who selected “none” for mobility, and positive for those who selected “mild” or “extreme” mobility issues. For White respondents, the biggest increase of the mean standardized score occurs between “Moderate” and “Severe”, while the change of the score from “Severe” to “Extreme” is much smaller. Compared to White respondents, for Black and Hispanic speaking Spanish, the change of the mean scores from “Moderate” to “Severe” is similar to change from “Severe” to “Extreme”. Note that for Hispanics speaking English, the mean score is lower among those who select “Extreme” compared to those who select “Moderate” or “Severe”, while for all other groups, the standardized score increases as the severity of the response categories increase. This indicates that respondents from different racial / ethnic groups use the scales differently, leading to DIF, and indicates the need to use methods like anchoring vignettes to achieve cross-cultural comparability.

7 We examined DIF across race/ethnicity and other socio-demographic groups, including gender, education and marital status. DIF were found across race/ethnicity groups but no other socio-demographic groups.

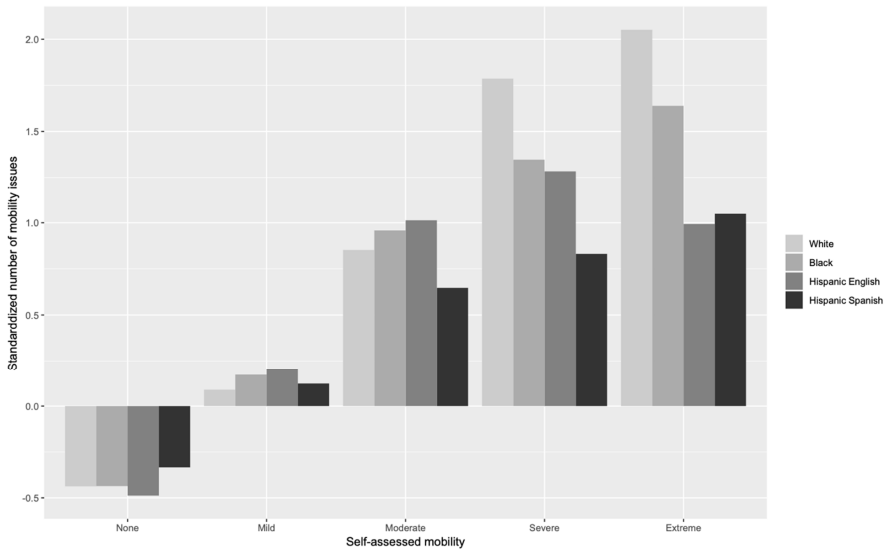


Figure 3 Mean standardized number of mobility issues by reported response categories of self-assessed mobility.

RQ 1. Response time

As shown in Table 2, regardless of domain, the average time respondents spent on a text vignette question is about twice as long as time spent on an image vignette question. Results for the statistical test of differential response time by vignette types using multilevel models are presented in Appendix 3 (Supplemental Table 6), which show consistent results as Table 2.

Table 2 Average time (in seconds) spent on one text or image vignette question by health domains.

	Pain		Sleep		Mobility		Affect	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Text vignette	15.93	8.81	15.73	8.25	17.85	10.31	18.05	10.59
Image vignette	7.95	3.58	8.38	3.61	8.33	3.79	7.42	3.17

RQ 2a. VE Test

Results of two tests of VE, the correct rank ordering test and the VE statistical test, were presented below.

Correct Rank-Ordering. Table 3 shows the percent of respondents whose ratings for the vignettes are consistent with the expected order (i.e., low intensity to high intensity). The percentages ranged from 17% to around 82%, depending on the domain. It is noted that for each of the four domains, the percentage of consistent rankings is significantly higher for the image than for the text vignette condition. In other words, respondents assigned to the image conditions are more likely to agree on the rank order of the vignettes than those assigned to the text condition. Respondents seem to have difficulty differentiating the rank orders of sleep and mobility *text vignettes*, with less than 20% able to correctly rank vignettes for these domains⁸. We also formally tested the effects of vignette types on the rank ordering of vignettes by fitting logistic regression models for each health domain (Results not shown). Not surprisingly, the odds of correctly ranking vignettes in the image vignette conditions are significantly higher compared to those in text vignette conditions. This is consistent across all four domains. Similar results were found when allowing for ties.

Statistical test of VE. Table 4 presents the results of statistical test of VE. The VE assumption is rejected in almost all conditions, except for the sleep text vignettes.

Table 3 Percentage of respondents ordering vignettes consistently with expected ordering.

	Pain		Sleep		Mobility		Affect	
	n	%	n	%	n	%	n	%
Text vignette	1051	47.6	1051	17.7	1051	19.8	1051	67.1
Image vignette	2098	79.7	2098	74.0	2098	43.4	2098	81.8

Table 4 Likelihood ratio tests of vignette equivalence.

	Pain		Sleep		Mobility		Affect	
	df	LR Test	df	LR Test	df	LR Test	df	LR Test
Text vignettes	24	70.4***	24	24.4	24	55.1***	24	110.9***
Image vignette	24	137.4***	24	158.8***	24	67.1***	24	154.3***

*: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

8 This analysis was also performed when two tie situations were allowed: 1) ties between the first two vignettes (low and moderate intensity) and 2) ties between the last two vignettes (moderate and high intensity). Results of rank order test allowing ties are consistent with Table 3.

Table 5 presents the results for predicting vignette locations (i.e., where it lies on the latent health spectrum) for both text and image vignette conditions of each domain⁹. In Table 5, Vignette 3 is the reference vignette, the one describing the highest pain level. Gender, marital status and racial/ethnic groups are the main predictors that drive the violations of VE for pain text vignettes. As for pain image vignettes, gender, age, income, and racial/ethnic groups are the main predictors that drive the violations of VE.

Those who are married view the first pain text vignette (the vignette with the least pain) as further away from the reference vignette on the latent spectrum, with a positive coefficient of 0.31 ($p = 0.02$). In other words, married respondents view the first pain text vignette as depicting better health (or less pain) than those who are not married. Males view the first pain text vignette as depicting worse health (or more pain) than females, which is consistent for both text and image AV conditions. Note that racial/ethnic group differences are significant for all health domains, suggesting that respondents from different racial/ethnic groups view the vignettes differently. For example, Hispanics interviewed in Spanish view Vignette 1 as depicting more pain than White respondents, regardless of text or image vignette designs.

As shown in Table 5, racial/ethnic group is a predictor that drives violations of VE for all health domains. To further examine this, Figure 4 presents the estimated vignette locations relative to the reference vignette by racial/ethnic group and vignette type for each health domain. If VE is satisfied, we would expect the estimated pain vignette locations to be exactly the same for each racial/ethnic group. This is not the case, as can be seen from Table 5 and Figure 4. As shown in Figures 4A1 and 4A2, Hispanics who completed the Spanish-language survey view the first vignette person (least severity) as having more pain (i.e., closer to 0 line, the reference vignette with the highest severity) compared to White respondents. On the other hand, Hispanics who completed the English-language survey also view the first vignette person as having more pain than do White respondents under the text condition, but not under the image vignette condition. Similar results are found for the affect domain (see Figure 4D1 and 4D2).

Figures 4B1 and 4B2 shows the estimated vignette locations for the sleep domain. As can be seen from Figure 4B1, the estimated vignette locations across racial/ethnic groups are very similar, indicating that respondents regardless of racial/ethnic background view the vignettes in similar ways. However, it is worth noting that the perceived vignette location for the second vignette is not significantly different from the reference vignette, suggesting that the sleep text vignettes failed to provide a good distinction between the second and third vignettes. As

9 As a sensitivity analysis, we also fit models combining image and text vignettes in one model for each domain (i.e., treating vignette type as a predictor in the model). Results (shown in Appendix 4) suggests image vignettes perform better in distinguishing the intensity levels of the three vignettes for each domain.

Table 5 Predictors for perceived vignette locations on the latent health spectrum.

	Pain		Sleep		Mobility		Affect	
	Text	Image	Text	Image	Text	Image	Text	Image
<i>Vignette 1 (no/mild difficulty/intensity)</i>								
Constant	3.20***	5.12***	1.48***	5.51***	2.03***	2.04***	4.01***	5.72***
Married	0.31*	0.24	0.02	-0.13	0.27	0.06	0.37**	0.23
Male	-0.49***	-0.36**	-0.29**	-0.15	-0.30**	-0.05	-0.58***	-0.23
Employed	-0.1	0.00	0.06	0.20	-0.05	0.18*	0.06	0.15
More than high school	-0.09	0.09	0.16	0.22	0.02	0.18**	0.10	-0.24
Age 18 - 29	0.25	-0.57*	0.14	-0.53*	-0.08	-0.08	-0.05	-0.90***
Age 30 - 49	-0.17	0.01	0.14	-0.10	-0.10	-0.06	0.07	-0.69**
Age 50 - 64	0.09	-0.50*	0.15	-0.29	0.08	-0.10	0.05	-0.88***
Middle income	0.04	-0.28*	0.05	-0.43**	0.03	-0.02	-0.24	-0.33*
High income	-0.15	-0.15	-0.02	-0.29	-0.34*	-0.32**	-0.61**	-0.41*
Black	-0.12	0.02	-0.17	-0.25	-0.13	-0.19	-0.65**	-0.51**
Hispanic (English)	-0.47**	0.04	-0.21	-0.38	-0.04	-0.09	-0.46*	0.00
Hispanic (Spanish)	-0.82***	-1.21***	-0.13	-1.56***	-0.52**	-0.50***	-1.46***	-1.34***
<i>Vignette 2 (moderate difficulty/intensity)</i>								
Constant	1.38***	1.84***	0.01	1.43***	-0.43*	0.97***	2.48**	1.88***
Married	0.09	0.01	0.03	-0.15	0.20*	-0.03	0.19	0.04
Male	-0.11	-0.19*	0.01	0.04	0.03	-0.03	-0.40**	0.02
Employed	-0.03	0.03	0.14	0.01	0.06	0.15*	0.15	0.07
More than high school	-0.12	0.03	0.02	0.05	-0.01	0.20**	-0.08	-0.09
Age 18 - 29	0.13	-0.20	0.11	0.14	0.01	0.10	-0.17	-0.32*
Age 30 - 49	-0.06	-0.11	0.12	0.24	0.02	-0.01	-0.22	-0.20
Age 50 - 64	0.01	-0.24	0.10	0.04	0.12	-0.19	-0.18	-0.22
Middle income	0.06	-0.09	-0.06	-0.16	0.09	-0.07	-0.18	-0.05
High income	-0.12	0.00	-0.06	-0.21*	-0.19	-0.18	-0.31	-0.12
Black	0.27	-0.05	0.03	-0.19	0.12	-0.22*	-0.25	-0.16
Hispanic (English)	-0.13	0.12	-0.03	-0.08	0.06	-0.11	-0.22	-0.13
Hispanic (Spanish)	0.02	-0.30**	0.02	-0.35**	0.20	-0.31**	-0.87***	-0.39***

Notes: Vignette 3 (highest difficulty/intensity) is the reference vignette. *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

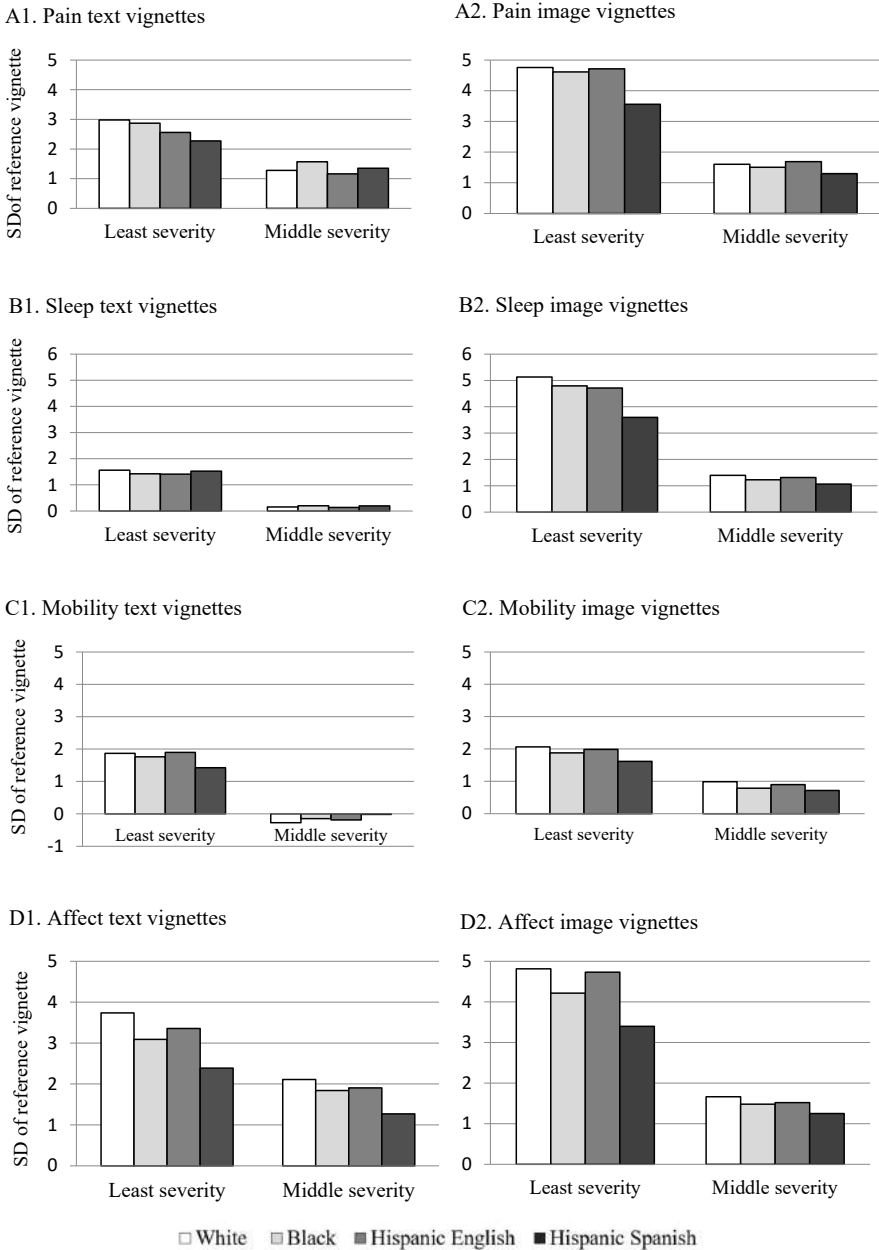


Figure 4 Estimated vignette locations, compared to the reference vignette (severity 3) on the latent health spectrum (measured in standard deviations of the reference vignette) for each health domain. Zero on the y-axis represents the mean of the reference (most pain or least healthy) vignette; higher numbers represent better perceived health.

shown in Figure 4B2, despite the VE violation (e.g., Hispanics who took the Spanish survey view the first vignette person as having more sleep difficulties compared with White respondents), image vignettes did a much better job differentiating the intensity levels of the three vignettes. Similar results are found for mobility domain (see Figures 4C1 and 4C2).

RQ 2b. RC-Test

As described in the *Analysis* section, the RC assumption test is based on visual comparisons of two sets of predicted mean cut points: one from Model A which has only vignettes (i.e., no self-assessments included in the model) and another from Model C which includes self-assessments and objective measures. Figure 5 shows the estimated cut points for all four health domains. If the vignette-derived cut point patterns are similar to the health measures-derived cut points, this indicates no or only minor violations of RC. For pain domain, both text and image vignettes show minor violations of RC. For all other three domains, image vignette conditions seem better fulfill RC, compared to text conditions.

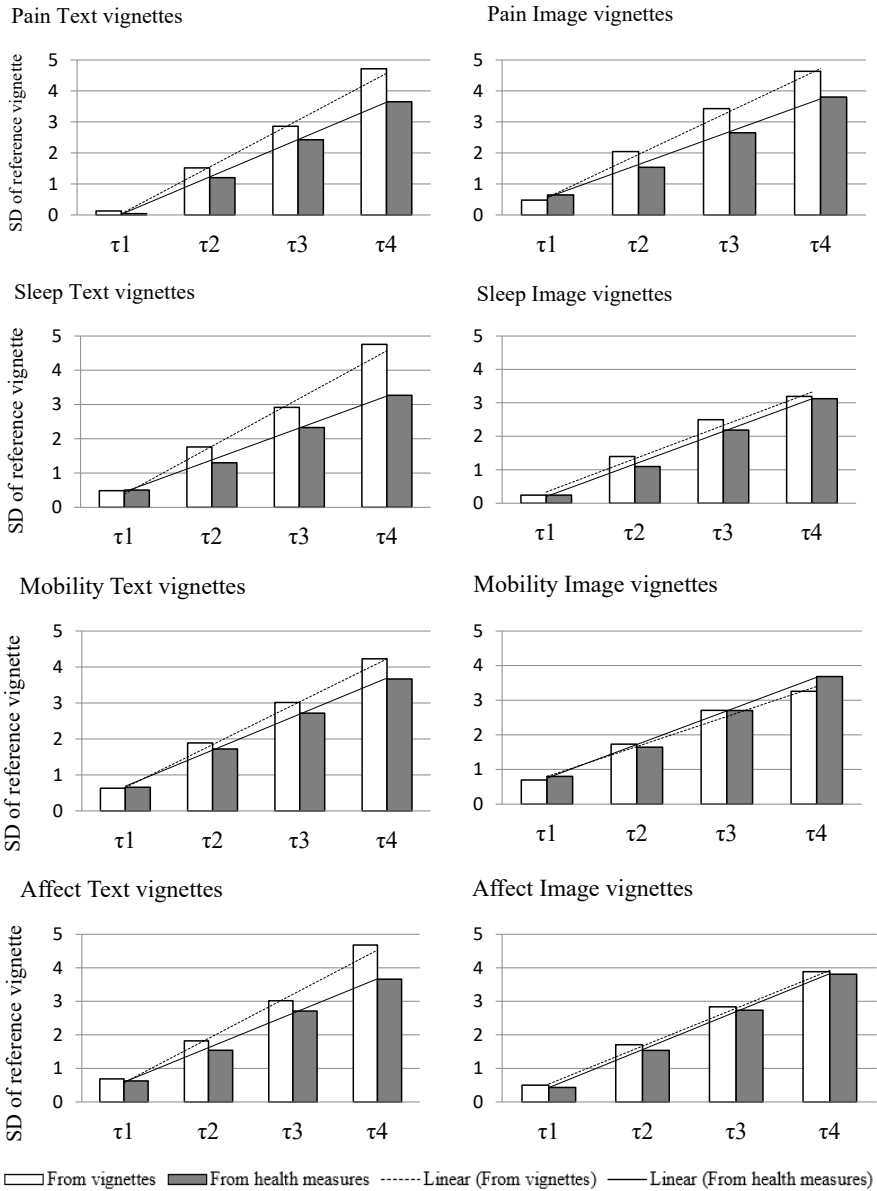


Figure 5 Estimated cut points for health domains based on vignettes and health measures. Evaluations are based on comparisons to the reference vignette [highest severity; measured in standard deviations (SD) of the reference vignette]. τ_1 – τ_4 are cut points for the five-point response scale from “None” to “Extreme” (e.g., τ_1 is the cut point between “None” and “Mild”).

RQ 3. Confirmatory factor analysis before and after anchoring vignette-adjustments

To test whether image and text AV-adjusted scores perform better than original scores (RQ 2a.), we compared model fit indices in a confirmatory factor analysis (CFA) using both adjusted scores and original scores. The cutoff criteria for acceptable fit are presented in the Analysis Strategy section. As shown in Table 6, CFI are above 0.95 and TLI are around or above 0.95 for all models, indicating that the models fit the data well for all the conditions. Models with AV-adjusted scores lead to better (i.e., higher) CFI and TLI values. For example, for the image condition subsample, the TLI of the model with image AV-adjusted self-assessment scores is 0.977, which is higher than the TLI of the model with original self-assessments – 0.942. RMSEA results shows that using both text and image AV-adjusted scores can greatly improve RMSEA. This suggests that using both text and image-adjusted scores improve CFA model fit.

To test whether image AV-adjusted scores perform similar or better than text AV-adjusted scores in the CFA framework (RQ 2b.), we assessed the model fit indices in CFA with text AV-adjusted scores and CFA with image AV-adjusted scores. As shown in Table 6, both text and image AV-adjustment improves the CFA based on original self-reports with similar improvements in terms of model fit indices. In addition, the CFI, TLI and RMSEA results are similar across the two CFA models with text vs. image AV-adjusted scores. The CFA with image AV-adjusted scores have better 90% CI of RMSEA (which ideally should have the lower value less than 0.05 and the upper value less than 0.08).

Table 6 Confirmatory Factor Analysis model fit estimates based on the original and anchoring vignette-adjusted scores.

Model	N	CFI	TLI	RMSEA	90% CI of RMSEA
CFA with original self-assessments (full sample)	3,149	0.983	0.948	0.158	(0.138, 0.179)
<i>Text condition subsample</i>					
CFA with original self-assessments	1,051	0.986	0.958	0.151	(0.117, 0.189)
CFA with text AV – adjusted self-assessment scores	1,051	0.994	0.982	0.060	(0.026, 0.100)
<i>Image condition subsample</i>					
CFA with original self-assessments	2,098	0.981	0.942	0.162	(0.137, 0.188)
CFA with image AV – adjusted self-assessment scores	2,098	0.992	0.977	0.062	(0.038, 0.089)

Discussion

This study examines the use of image anchoring vignettes (AV) to adjust DIF in self-assessments of health. Despite the fact that text AV have been adopted in many comparative studies, there are several critical challenges associated with text AV. To explore ways to overcome these challenges, this paper proposes the use of image AV, consisting of carefully designed and pre-tested images. In this study, the performances of text and image AV are compared with respect to a number of properties, including response time, tests of assumptions, and CFA model fits. Overall, the results suggest that the image AV methodology can be used as an improved and effective alternative to text AV in cross-cultural research, although the extent to which the VE assumption is satisfied needs further investigation for both text and image AV.

Specifically, the use of image AV can reduce survey time to about half the time of text AV. This result is consistent with previous literature on differences of information processing between text vs. image stimuli (Azizian et al., 2006; Naspetti et al. 2016; Schlochtermeyer et al. 2013). Survey time is an important indicator for respondent cognitive burden, which can influence survey data quality and survey response rates. Survey time is also closely associated with survey cost, with shorter time potentially implying lower survey costs. Thus, image AV offers a time and potentially cost-efficient survey option, compared to text AV, especially in studies with many AV items (e.g., Weiss & Roberts, 2018).

Results for comparing how well AV assumptions are satisfied between text and image AV show mixed findings. On the one hand, image AV outperforms text AV in that respondents can better distinguish the different intensity levels in image vignettes (e.g., from no pain to extreme pain) than in text vignettes, indicating that respondents are more likely to perceive the vignettes in similar ways and in the designed order in the image AV condition compared to the text AV condition. This finding is consistent with previous literature showing the information processing advantage of emotional images in terms of larger or more pronounced emotion effects evoked by image stimuli, compared to text stimuli (e.g., Schlochtermeyer et al., 2013). One of the reasons may be that image vignettes lead to a stronger activation of relevant information in the cognitive system resulting in more arousal and perceived intensity. Another possible reason is that text AV puts a higher cognitive burden on respondents, potentially resulting in more satisficing behavior including straight-lining (i.e., respondents select the same response option for all the vignette questions) and random selection of responses. For example, we find that respondents assigned to the text vignettes treatment are more likely to straight-line than those assigned to image vignettes (results not shown).

On the other hand, for both text and image AV, it is found that respondents' perceptions of the vignettes can differ by cultural subgroups, a violation of VE.

Similar to text AV, various factors may cause violations of VE for image vignettes. First, like text vignettes, the information in image AV may serve as memory cues that can trigger other related memories, leading to differences in perceptions. Second, although elements included in image AV may be more easily standardized than text AV (e.g., gender of the hypothetical person), the included elements may still weigh differently for different subgroups. For example, an element in the image may be more familiar to one cultural group than to another, resulting in perception differences. The violation of VE implies that designing “universal” anchoring vignettes (Grol-Prokopczyk, 2018), which are familiar to all population groups and reveal the same information to all respondents, is still a challenge for both text and image vignettes.

Despite the VE violations, results of the CFA models indicate that, compared to the model with self-reported data, using vignettes-adjusted scores can greatly improve model fit, which is consistent with Weiss & Roberts (2018)¹⁰. This shows that, even though VE is not met, it is still better to use text or image AV-adjustments, which can effectively reduce measurement errors. Comparing the two vignette types, text and image vignettes perform similarly in terms of measurement error reduction in the CFA models.

Given the clear advantage of image vignettes in reducing survey time, lowering respondents’ cognitive burden and better differentiating intensity levels, we believe there is a potential for the use of image AV to improve text AV methodology.

This study also revealed important findings to deepen our understanding of the vignette methodology, including how different respondents view and rate vignettes. For example, it was found that male respondents view the first pain vignette as describing more pain than female respondents do (as shown in Table 5). This may be because females experience more pain than males (Cepeda & Carr, 2003). They may use themselves as a standard of comparison when rating the vignette person and thus view the first vignette person as depicting minimal pain. Due to space restrictions, this study will not discuss detailed results for all covariates. Future studies can look into this further. In addition, this study generates new insights into the design and use of image AV, and the designed image AV items can be applied to other studies that use anchoring vignettes to adjust self-reported health.

It is worth mentioning that this study is limited in several ways. First, due to resource constraints, our experimental study is based on a non-probability sample, from which the results were not intended to generalize to the full U.S. population. Among the four types of validity of causal inference (statistical, internal, external and construct validity) in Shadish, Cook and Campbell (2002), this paper focused

10 We also examined the DIF-adjusting results using HOPIT models. Results are similar for both text and image vignettes. Due to space restraints, results are not shown in this paper and are available upon request.

on the internal and statistical validity with a randomized experiment to compare DIF-adjustment results between vignette types. Per Edgington (1966) and Berk et al. (1995), randomized experiments permit statistical inferences about the experimental factors. However, due to the nature of the sample, we do not claim that our results generalize to the complete U.S. population and beyond. Future studies could replicate this study in probability-based representative surveys to evaluate the effect sizes of the group comparisons in the population. Second, the current RC test is not based on a statistical test and additional evaluations of RC using more stringent RC test are needed (Grol-Prokopczyk, 2018). Third, the objective health measures used in the RC tests may not fully capture actual health. One may also argue that these objective health questions are based on self-reports and may be subject to reporting errors. Note that the questions about objective health are straightforward factual questions (e.g., whether respondent has received doctor diagnosis of certain diseases), for which reporting errors may be less of an issue compared to self-assessing of a health domain. Also, many of the objective measures used in this study are based on widely-used existing scales, and have been successfully applied in previous literature (Kessler et al., 2002; Ray et al., 2009). If available, future studies could use bio-markers (e.g., medical test results and genetic data) in the RC tests. Fourth, this study examined the most commonly used text vignettes that are included in HRS, SHARE, and many other large-scale surveys. It is possible that text vignettes with differently-worded descriptions may perform better in tests of assumptions than the current text vignettes. The same may be true for image vignettes. Possibly, better-designed pictures are less likely to lead to rejection of the VE and RC assumptions. Future research could compare text and image vignettes with different descriptions or designs.

Our research suggests several important directions for future research. First, this study focuses on the comparisons of text and image vignettes in correcting for DIF. Future research could examine in detail how different image vignette designs may influence the performance of image AV. For example, in a related study, we found that when rating image vignettes with average body size vs. obese for the mobility domain, respondents tend to rate the obese vignette person as having more mobility difficulties than a vignette person with average body size. This is not surprising given that obese individuals are more likely to have mobility limitations than non-obese individuals (Koster et al., 2007). In addition, the vignette images showing average body sizes, which match the body size of the majority of respondents, show a higher rate of consistency in the rank-orderings, indicating that respondents may better perceive the image vignettes when the vignette figures match more closely their own characteristics. This could shed light on the future design of image vignettes. For example, it indicates that image vignettes that have a broader applicability and familiarity to the respondents may better satisfy the assumptions. Future research could further evaluate the effects of a wide range of

vignette characteristics on image vignette performance. Second, given budget constraints, all respondents in this study are from the U.S. Future research could evaluate the use of vignettes in a less homogeneous group, such as extending the study to cross-national surveys and/or to a wide variety of other racial/ethnic groups, such as Asians, American Indian or Alaska Native, and Native Hawaiian or Other Pacific Islander. Third, some domains may be too complex to be expressed using images, such as self-reported political attitudes. In addition, using static image vignettes may not be the best way to present measures related to change over time and location, such as a slow or fast walking speed. Future research can evaluate other visual vignette designs such as using short videos in web surveys (Banuri et al., 2018; Mendelson, Gibson, & Romano-Bergstrom, 2017) and the use of visual vignettes in different domains, including domains that cannot be easily visualized using static images. Fourth, the ways vignettes are presented and their applications can vary by survey mode, which may influence their performance. Verbal vignettes can be delivered orally in telephone and face-to-face interviews or visually as text in mail and web surveys, but image vignettes have to be presented visually in mail and web surveys, or as a picture presented by interviewers in face-to-face surveys. Future research could evaluate mode effects for both text and image vignettes.

In conclusion, this study indicates that using either text or image AV adjustments can reduce measurement errors compared to the analysis without using any AV, and the use of image AV can greatly reduce survey time and respondents' cognitive burden as compared to text vignettes. Improving VE, (in other words, minimizing different interpretations of vignettes by different groups), is critical for both text and image AV and requires further investigation. This study has advanced knowledge of the design and applications of image AV in health surveys and has implications for designing image AV of other domains. Future implementations of AV can use the findings of this study to introduce efficiencies in their survey designs.

References

- Azizian, A., Watson, T. D., Parvaz, M. A., & Squires, N. K. (2006). Time course of processes underlying picture and word evaluation: an event-related potential approach. *Brain Topography*, 18(3), 213-222.
- Banuri, S., de Walque, D., Keefer, P., Haidara, O. D., Robyn, P. J., & Ye, M. (2018). The use of video vignettes to measure health worker knowledge. Evidence from Burkina Faso. *Social Science & Medicine*, 213, 173-180.
- Berk, R. A., Western, B., & Weiss, R. E. (1995). Statistical inference for apparent populations. *Sociological methodology*, 421-458.
- Bolt, D. M., Lu, Y., & Kim, J.-S. (2014). Measurement and control of response styles using anchoring vignettes: A model-based approach. *Psychological Methods*, 19(4), 528-541.

- Buskirk, T. D. (2015). Are sliders too slick for surveys? An experiment comparing slider and radio button scales for smartphone, tablet and computer based surveys. *methods, data, analyses*, 9(2), 32.
- Caramazza, A. (1996). Pictures, words and the brain. *Nature*, 383(6597), 216-217.
- Cepeda, M. S., & Carr, D. B. (2003). Women experience more pain and require more morphine than men to achieve a similar degree of analgesia. *Anesthesia and Analgesia*, 1464-1468.
- Couper, M. P. (2005). Technology trends in survey data collection. *Social Science Computer Review*, 23(4), 486-501.
- Couper, M. P., Conrad, F. G., & Tourangeau, R. (2007). Visual context effects in web surveys. *Public Opinion Quarterly*, 71(4), 623-634.
- Couper, M. P., Tourangeau, R., Conrad, F. G., & Singer, E. (2006). Evaluating the effectiveness of visual analog scales: A web experiment. *Social Science Computer Review*, 24(2), 227-245.
- Couper, M. P., Tourangeau, R., & Kenyon, K. (2004). Picture this! Exploring visual effects in web surveys. *Public Opinion Quarterly*, 68(2), 255-266.
- d'Uva, T. B., Lindeboom, M., O'Donnell, O., & Van Doorslaer, E. (2011). Slipping anchor? Testing the vignettes approach to identification and correction of reporting heterogeneity. *Journal of Human Resources*, 46(4), 875-906.
- d'Uva, T., O'Donnell, O., & Van Doorslaer, E. (2008). Differential health reporting by education level and its impact on the measurement of health inequalities among older Europeans. *International Journal of Epidemiology*, 37(6), 1375-1383.
- Davis, R. E., Johnson, T. P., Lee, S., & Werner, C. (2019). Why do Latino survey respondents acquiesce? Respondent and interviewer characteristics as determinants of cultural patterns of acquiescence among Latino survey respondents. *Cross-Cultural Research*, 53(1), 87-115.
- DeSalvo, K. B., Bloser, N., Reynolds, K., He, J., & Muntner, P. (2006). Mortality prediction with a single general self-rated health question. *Journal of general internal medicine*, 21(3), 267-275
- Dowd, J. B., & Todd, M. (2011). Does self-reported health bias the measurement of health inequalities in US adults? Evidence using anchoring vignettes from the Health and Retirement Study. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 66(4), 478-489.
- Edgington, E. S. (1966). Statistical inference and nonrandom samples. *Psychological Bulletin*, 66(6), 485.
- Ferrer-i-Carbonell, A., Van Praag, B. M. S., & Theodossiou, I. (2011). Vignette Equivalence and Response Consistency: The Case of Job Satisfaction. Retrieved from <http://papers.ssrn.com/abstract=1968870>
- Glaser W. R. (1992). Picture naming. *Cognition*: 42(1-3), 61-105.
- Glaser W. R., Glaser M. O. (1989). Context Effects in Stroop-Like Word and Picture Processing. *Journal of Experimental Psychology: General*: 118(1), 13-42.
- Grady, C. L., McIntosh, A. R., Rajah, M. N., & Craik, F. I. (1998). Neural correlates of the episodic encoding of pictures and words. *Proceedings of the National Academy of Sciences*, 95(5), 2703-2708.
- Gravelle, T. B. (2021). The Measurement Invariance of Customer Loyalty and Customer Experience across Firms, Industries, and Countries. *methods, data, analyses*, 23

- Grol-Prokopczyk, H. (2018). In pursuit of anchoring vignettes that work: Evaluating generality versus specificity in vignette texts. *The Journals of Gerontology: Series B*, 73(1), 54-63.
- Grol-Prokopczyk, H., Verdes-Tennant, E., McEniry, M., & Ispány, M. (2015). Promises and pitfalls of anchoring vignettes in health survey research. *Demography*, 52(5), 1703-1728.
- Groot, T. D., Jacquet, W., Backer, F. D., Peters, R., & Meurs, P. (2020). Using image vignettes to explore sensitive topics: a research note on exploring attitudes towards people with albinism in Tanzania. *International Journal of Social Research Methodology*, 1-7.
- Hibbing, A. N., & Rankin-Erickson, J. L. (2003). A picture is worth a thousand words: Using visual images to improve comprehension for middle school struggling readers. *The reading teacher*, 56(8), 758-770.
- Hirve, S., Gomez-Olive, X., Oti, S., Debuur, C., Juvekar, S., Tollman, S., ... & Ng, N. (2013). Use of anchoring vignettes to evaluate health reporting behavior amongst adults aged 50 years and above in Africa and Asia—testing assumptions. *Global health action*, 6(1), 21064.
- Holt, T. P., & Loraas, T. M. (2019). Using Qualtrics panels to source external auditors: A replication study. *Journal of Information Systems*, 33(1), 29-41.
- Hopkins, D. J., & King, G. (2010). Improving anchoring vignettes: Designing surveys to correct interpersonal incomparability. *Public opinion quarterly*, 74(2), 201-222.
- Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1-55.
- Hu, M., & Lee, S. (2016). Context Effects in Anchoring Vignette Questions. The 71st Annual Conference of the American Association for Public Opinion Research, Austin, Texas.
- Hu, M., Lee, S., & Xu, H. (2018). Using Anchoring Vignettes to Correct for Differential Response Scale Usage in 3MC Surveys. In T. P. Johnson, B.-E. Pennell, I. Stoop, & B. Dorer (Eds.), *Advances in Comparative Survey Methodology*.
- Ibarra, J. L., Agas, J. M., Lee, M., Pan, J. L., & Buttenheim, A. M. (2018). Comparison of online survey recruitment platforms for hard-to-reach pregnant smoking populations: feasibility study. *JMIR research protocols*, 7(4), e8071.
- Idler, E. L., & Benyamini, Y. (1997). Self-rated health and mortality: a review of twenty-seven community studies. *Journal of health and social behavior*, 21-37.
- Idler, E. L., & Kasl, S. V. (1995). Self-ratings of health: do they also predict change in functional ability?. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 50(6), S344-S353.
- Jürges, H., & Winter, J. (2013). Are anchoring vignettes ratings sensitive to vignette age and sex? *Health Economics*, 19(22), 1-13.
- Kapteyn, A., Smith, J. P., Van Soest, A., & Vonková, H. (2011). Anchoring Vignettes and Response Consistency Consistency. *Working Paper*. Retrieved from http://www.rand.org/content/dam/rand/pubs/working_papers/2011/RAND_WR840.pdf
- Kessler, R. C., Andrews, G., Colpe, L. J., Hiripi, E., Mroczek, D. K., Normand, S. L., ... & Zaslavsky, A. M. (2002). Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychological medicine*, 32(6), 959-976.

- King, G., Murray, C. J. L., Salomon, J. A., & Tandon, A. (2004). Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research. *American Political Science Review*, 98, 191–207.
- King, G., & Wand, J. (2007). Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes. *Political Analysis*, 15, 46–66.
- Koster, A., Penninx, B. W. J. H., Newman, A. B., Visser, M., Van Gool, C. H., Harris, T. B., ... Kritchevsky, S. B. (2007). Lifestyle factors and incident mobility limitation in obese and non-obese older adults. *Obesity*, 15(12), 3122–3132.
- Kristensen, N., & Johansson, E. (2008). New evidence on cross-country differences in job satisfaction using anchoring vignettes. *Labour Economics*, 15(1), 96–117.
- Krosnick, J. A., & Abelson, R. P. (1992). The case for measuring attitude strength in surveys. In *Questions About Questions: Inquiries into the Cognitive Bases of Surveys* (pp. 177–203). Russell Sage Foundation. Retrieved from https://books.google.com/books?hl=en&lr=&id=8FEiM0gA_wwC&pgis=1
- Kyllonen, P. C., & Bertling, J. P. (2014). Anchoring vignettes reduce Bias in noncognitive rating scale responses. *Report Submitted to OECD*.
- Lee, J. W., Jones, P. S., Mineyama, Y., & Zhang, X. E. (2002). Cultural differences in responses to a Likert scale. *Research in Nursing and Health*, 25, 295–306.
- Lee, S., Liu, M., & Hu, M. (2017). Relationship between future time orientation and item nonresponse on subjective probability questions: A cross-cultural analysis. *Journal of cross-cultural psychology*, 48(5), 698-717.
- Lee, S., Schwarz, N., & Goldstein, L. S. (2014). Culture-sensitive question order effects of self-rated health between older Hispanic and non-Hispanic adults in the United States. *Journal of aging and health*, 26(5), 860-883.
- Liu, M., Kuriakose, N., Cohen, J., & Cho, S. (2016). Impact of web survey invitation design on survey participation, respondents, and survey responses. *Social Science Computer Review*, 34(5), 631–644.
- Luna, D., & Peracchio, L. A. (2003). Visual and linguistic processing of ads by bilingual consumers. *Persuasive Imagery: A Consumer Response Perspective*, 153–175.
- MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological methods*, 1(2), 130.
- McCarthy, M., Ruiz, E., Gale, B., Karam, C., & Moore, N. (2004). The meaning of health: Perspectives of Anglo and Latino older women. *Health Care for Women International*, 25(10), 950-969
- Mendelson, J., Gibson, J. L., & Romano-Bergstrom, J. (2017). Displaying Videos in Web Surveys: Implications for Complete Viewing and Survey Responses. *Social Science Computer Review*, 35(5), 654–665.
- Mojtabai, R. (2015). Depressed Mood in Middle-Aged and Older Adults in Europe and the United States: A Comparative Study Using Anchoring Vignettes. *Journal of Aging and Health*, 1–23.
- Mohler, P., Dorer, B., de Jong, J., & Hu, M. (2016). Translation: overview. *Guidelines for Best Practice in Cross-Cultural Surveys*. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan.
- Molina, T. (2016). Reporting Heterogeneity and Health Disparities Across Gender and Education Levels: Evidence From Four Countries. *Demography*, 53(2), 295–323.

- Murray, C. J. L., Tandon, A., Salomon, J. A., Mathers, C. D., & Sadana, R. (2002). New approaches to enhance cross-population comparability of survey results. *Summary Measures of Population Health: Concepts, Ethics, Measurement, and Applications*, 421–432.
- Murray, C. J., Ozaltin, E., Tandon, A., Salomon, J., Sadana, R., & Chatterji, S. (2003). Empirical evaluation of the anchoring vignettes approach in health surveys. In *Health systems performance assessment: Debates, methods and empiricism* (pp. 369–399).
- Naspetti, S., Mandolesi, S., & Zanoli, R. (2016). Using visual Q sorting to determine the impact of photovoltaic applications on the landscape. *Land Use Policy*, 57, 564–573.
- Naylor, R., Maye, D., Ilbery, B., Enticott, G., & Kirwan, J. (2014). Researching controversial and sensitive issues: using image vignettes to explore farmers' attitudes towards the control of bovine tuberculosis in England. *Area*, 46(3), 285–293.
- Paivio, A. (2013). *Imagery and verbal processes*. Psychology Press.
- Pan, Y., & De La Puente, M. (2005). Census Bureau guideline for the translation of data collection instruments and supporting materials: Documentation on how the guideline was developed. *Survey Methodology*, 6.
- Primi, R., Zanon, C., Santos, D., De Fruyt, F., & John, O. P. (2016). Anchoring vignettes can they make adolescent self-reports of social-emotional skills more reliable, discriminant, and criterion-valid? *European Journal of Psychological Assessment*, 32(1), 39–51.
- Ray, L., Lipton, R. B., Zimmerman, M. E., Katz, M. J., & Derby, C. A. (2009). Mechanisms of association between obesity and chronic pain in the elderly. *Pain*.
- Revilla, M. (2017). Analyzing survey characteristics, participation, and evaluation across 186 surveys in an online opt-in panel in Spain. *methods, data, analyses*, 11(2), 28.
- Rice, N., Robone, S., & Smith, P. C. (2012). Vignettes and health systems responsiveness in cross-country comparative analyses. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 175(2), 337–369.
- Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of educational research*, 99(6), 323–338.
- Sevostianov, A., Horwitz, B., Nechaev, V., Williams, R., Fromm, S., & Braun, A. R. (2002). fMRI study comparing names versus pictures of objects. *Human brain mapping*, 16(3), 168–175
- Schlochtermeyer, L. H., Kuchinke, L., Pehrs, C., Urton, K., Kappelhoff, H., & Jacobs, A. M. (2013). Emotional picture and word processing: an fMRI study on effects of stimulus complexity. *PLoS One*, 8(2), e55619
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin
- Townsend, C., & Kahn, B. E. (2014). The “Visual Preference Heuristic”: The Influence of Visual versus Verbal Depiction on Assortment Processing, Perceived Variety, and Choice Overload. *Journal of Consumer Research*, 40(5), 993–1015.
- Van Soest, A., Delaney, L., Harmon, C., Kapteyn, A., & Smith, J. P. (2011). Validating the use of anchoring vignettes for the correction of response scale differences in subjective questions. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 174, 575–595.
- Wand, J., King, G., & Lau, O. (2011). Anchors: Software for anchoring vignette data. *Journal of Statistical Software*, 42(1), 1–25.







- Weiss, S., & Roberts, R. D. (2018). Using anchoring vignettes to adjust self-reported personality: A comparison between countries. *Frontiers in Psychology*, 9(MAR), 1–17.
- Witte, J. C., Pargas, R. P., Mobley, C., & Hawdon, J. (2004). Instrument effects of images in web surveys: A research note. *Social Science Computer Review*, 22(3), 363–369.
- Yan, T., & Hu, M. (2018). Examining Translation and Respondents' Use of Response Scales in 3MC Surveys. *Advances in Comparative Survey Methods*, 501–518.

APPENDIX 1







Text and image vignettes used for the web survey for each domain.

Note that in the design of image vignettes, we have two different design conditions per domain. Given that the aim of this paper is to compare text vs. image vignettes, data from different designs of image vignettes are combined in all the analysis. The evaluation the design of features on AV methodology is discussed elsewhere.

Supplemental Table 1 Pain text and image vignettes.

Pain Intensity Level	Text vignette	Image Design One (young adults)	Image Design Two (seniors)
No / Low Pain	Karen has a headache once a month that is relieved after taking a pill. During the headache she can carry on with her day-to-day affairs.		
Moderate Pain	Jennifer has pain that radiates down her right arm and wrist during her day at work. This is slightly relieved in the evenings when she is no longer working on her computer.		
High Pain	Mary has pain in her knees, elbows, wrists and fingers, and the pain is present almost all the time. Although medication helps, she feels uncomfortable when moving around, holding and lifting things.		










Supplemental Table 2 Sleep text and image vignettes.

Sleep Difficulty Level	Text vignette	Image Design One (female)	Image Design Two (male)
No / Low Difficulty	Sara/Sam falls asleep easily at night, but two nights a week she/he wakes up in the middle of the night and cannot go back to sleep for the rest of the night.		
Moderate Difficulty	Susan/Scott wakes up almost once every hour during the night. When she/he wakes up in the night, it takes around 15 minutes for him/her to go back to sleep. In the morning she/he does not feel well-rested.		
High Difficulty	Patty/Paul takes about two hours every night to fall asleep. She/He wakes up once or twice a night feeling panicked and takes more than one hour to fall asleep again.		

Supplemental Table 3 Mobility text and image vignettes.

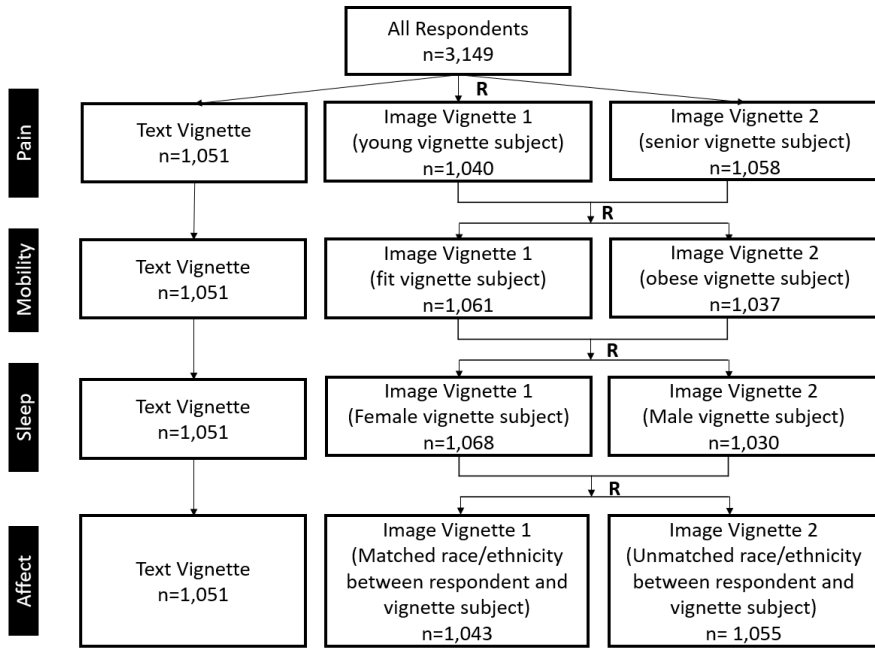
Mobility Difficulty Level	Text vignette	Image Design One (optimal weight/fit)	Image Design Two (obese)
No / Low Difficulty	Laura is able to walk distances of up to 200 metres without any problems but feels tired after walking one kilometre or climbing more than one flight of stairs. She has no problems with day-to-day activities, such as carrying food from the market.		
Moderate Difficulty	Sandy does not exercise. She cannot climb stairs or do other physical activities because she is obese. She is able to carry the groceries and do some light household work.		
High Difficulty	Lisa has a lot of swelling in her legs due to her health condition. She has to make an effort to walk around her home as her legs feel heavy.		

Supplemental Table 4 Affect text and image vignettes.

Depression Level	Text vignette	White	Black	Hispanic
No / Low Depression	Matt enjoys his work and social activities and is generally satisfied with his life. He gets depressed every 3 weeks for a day or two and loses interest in what he usually enjoys but is able to carry on with his day-to-day activities.			
Moderate Depression	David feels nervous and anxious. He worries and thinks negatively about the future but feels better in the company of people or when doing something that really interests him. When he is alone he tends to feel useless and empty.			
High Depression	Leo feels depressed most of the time. He weeps frequently and feels hopeless about the future. He feels that he has become a burden to others and that he would be better off dead.			

APPENDIX 2

Randomization conditions and assignments and robustness checks for randomization across text and image conditions.



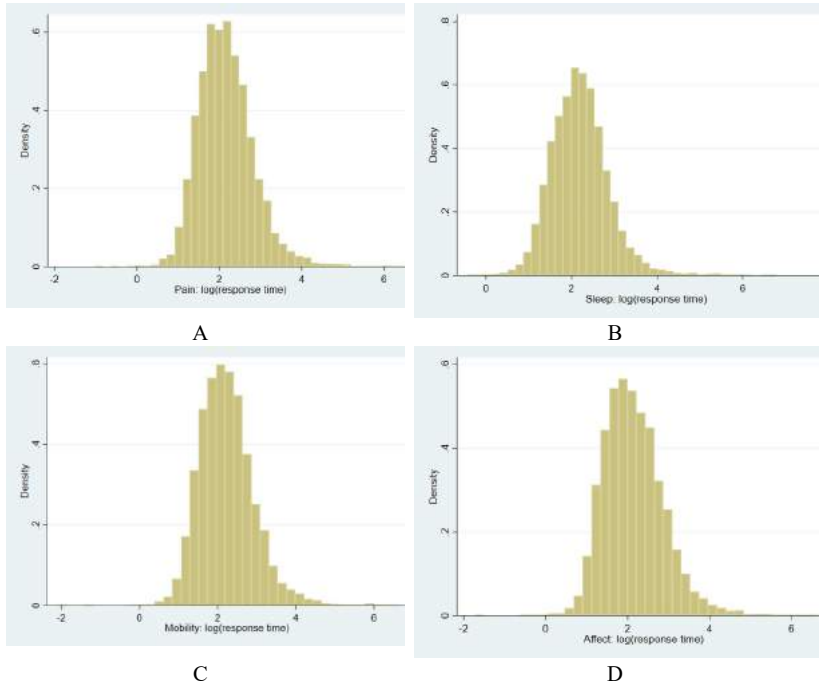
Supplemental Figure 1 Experimental conditions and assignments for each domain. “R” indicates randomization was done.

Supplemental Table 5 Robustness checks for randomization across text and image conditions.

	Text	Image	Chi-square / F statistics
Gender			0.03
Female	52.3	51.9	
Male	47.7	48.1	
Age (mean)	46.9	46.7	0.10
Race			0.47
White	23.8	24.3	
Black	23.8	23.9	
Non-Hispanic White	23.5	24.0	
Non-Hispanic Black	28.9	27.8	
Education			0.01
Below high school	52.2	52.0	
High school and above	47.8	48.0	
Employment status			1.40
Employed	52.6	54.9	
Not employed	47.4	45.1	
Marital status			0.55
Married	50.2	48.8	
Not married	49.8	51.2	
Income			0.84
Low	34.3	34.9	
Middle	42.2	40.6	
High	23.5	24.5	

APPENDIX 3

Distributions of log-transformed response time variable for each domain.



Supplemental Figure 2 Distributions of log-transformed response time variable for each domain.

To formally test the differential response time by vignette types, for each health domain, we fit multilevel logistic regression models with random intercepts. Given that time is right skewed, we used log-transformed time as outcomes (distributions shown in Appendix 3). In the unconditional model (i.e., no predictors in the model) for each domain, log-transformed response time varied significantly across individuals (the intraclass correlation coefficient [ICC] ranges from 0.43 to 0.50, see Supplemental Table 6), justifying the use of multilevel modeling. Supplemental Table 6 shows the results of the final models which include both question level predictors (i.e., image vs. text vignettes) and respondent level predictors (e.g., demographic and socio-economic variables). As shown in Supplemental Table 6, compared to text vignettes, respondents spent significantly less time answering image vignettes. This is true for all four domains. Compared to non-Hispanic White, respondents

of all other three groups spent significantly longer time in answering the vignette questions.

Supplemental Table 6 Multilevel linear regression models predicting log-transformed response time for each health domain.

	Model			
	Pain	Sleep	Mobility	Affect
Image vignettes (ref: Text vignettes)	-0.63***	-0.58***	-0.68***	-0.78***
Age	0.01***	0.01***	0.01***	0.01***
Male (ref: Female)	-0.01	0.00	-0.06**	-0.04*
Above high school education (ref: Below high school)	-0.05**	-0.30	-0.05*	-0.04*
Employed (ref: Not employed)	-0.06**	-0.60**	-0.04*	-0.02
Married (ref: Not married)	-0.05**	-0.45*	-0.04	-0.06**
Respondent Groups (Ref: Non-Hispanic White)				
Non-Hispanic Black	0.18***	0.20***	0.18***	0.18***
Hispanics English	0.06*	0.07**	0.09**	0.07**
Hispanics Spanish	0.16***	0.16***	0.14***	0.17***
ICC	0.50	0.45	0.49	0.43
(95% confidence interval)	(0.48, 0.52)	(0.43, 0.47)	(0.47, 0.51)	(0.41, 0.45)

*: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

APPENDIX 4

Model results for evaluating VE test for each domain (with both image and text vignettes combined for analysis).

Supplemental Table 7 Predictors for perceived vignette locations on the latent health spectrum.

	Pain	Sleep	Mobility	Affect
<i>Vignette 1 (no/mild difficulty/intensity)</i>				
Constant	3.39***	1.74***	1.90***	4.10***
Image	1.59***	2.84***	0.17**	1.07***
Married	0.21*	-0.07	0.12	0.15
Male	-0.43***	-0.23**	-0.14*	-0.24**
Employed	-0.02	0.15	0.10	0.12
More than high school	-0.03	0.11	0.12	-0.10
Age 18 - 29	-0.20	0.02	-0.12	-0.50**
Age 30 - 49	-0.21	0.10	-0.09	-0.40**
Age 50 - 64	-0.27*	0.09	-0.05	-0.39**
Middle income	-0.14	-0.11	0.00	-0.19*
High income	-0.05	-0.19	-0.31***	-0.46***
Black	-0.08	-0.13	-0.15	-0.44***
Hispanic (English)	-0.17	-0.14	-0.04	-0.17
Hispanic (Spanish)	-0.89***	-0.55***	-0.48***	-1.13***
<i>Vignette 2 (moderate difficulty/intensity)</i>				
Constant	1.58***	0.20	-0.14	2.21***
Image	0.16*	1.03***	0.98***	-0.27***
Married	0.01	-0.07	0.06	0.06
Male	-0.16*	0.00	0.00	-0.08
Employed	0.00	0.05	0.10	0.09
More than high school	-0.04	0.04	0.12*	-0.10
Age 18 - 29	-0.06	0.15	0.07	-0.33**
Age 30 - 49	-0.12	0.18	-0.01	-0.26*
Age 50 - 64	-0.15	0.07	-0.08	-0.24*
Middle income	-0.04	-0.08	-0.03	-0.10
High income	-0.05	-0.14	-0.18*	-0.18*
Black	0.05	-0.07	-0.12	-0.11
Hispanic (English)	0.00	-0.05	-0.05	-0.11
Hispanic (Spanish)	-0.15	-0.19*	-0.15*	-0.46***

Notes: Vignette 3 (highest difficulty/intensity) is the reference vignette. *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

New Methodical Findings on D-Efficient Factorial Survey Designs: Impacts of Design Resolution on Aliasing and Sample Size

Julia Kleinewiese

*Mannheim Centre for European Social Research (MZES),
University of Mannheim*

Abstract

In empirical surveys, finding a sufficient number of respondents can be challenging. For factorial survey experiments, drawing a vignette-sample (“fraction”) from a vignette-universe can reduce the minimum number of respondents required. Vignette-samples can be drawn by applying D-efficient designs. Theoretically, D-efficient resolution V designs are ideal. Due to reasons of practicability, however, resolution IV designs have usually been applied in empirical social research and are considered to be sufficient when it is clear up front, which two-way interactions are likely to have an effect. Against this backdrop, this article focusses on two research questions: (1) In resolution IV designs, are those two-way interactions that are not orthogonalized truly not aliased with any main effects? (2) How does design resolution affect the minimum size of the vignette-sample that is necessary for achieving an adequate level of D-efficiency? These questions are examined by applying SAS-macros for computing D-efficient samples, pre-construction assessment and post-construction evaluation. The resulting aliasing structures indicate a discrepancy between previous definitions of design resolutions and the aliasing structures of designs resulting from the SAS-macros. Additionally, they suggest taking a second look at the assumption that higher resolutions or larger vignette universes will always necessitate designs with larger vignette-samples (and thus larger sets or more respondents).

Keywords: D-efficiency, design resolution, sample size, factorial survey, aliasing, confounding



When collecting quantitative data, a major issue is finding a sufficient number of respondents (cf. Engel & Schmidt, 2019). Factorial surveys offer some unique opportunities based on design, such as reducing the number of required respondents, but also challenges that need to be considered carefully – from the design-stage onwards. Factorial survey vignettes are an established method in quantitative social science research measuring attitudes or behaviour. Methodical research implies that vignettes can have a high external validity, i.e. are suitable for measuring real-life attitudes and behaviour (Hainmueller et al., 2015) but may sometimes run into issues, such as social desirability bias (Eifler, 2007; Eifler et al., 2014). Nevertheless, they are especially important for research that cannot be conducted in real-life, due to practical or ethical considerations (such as research on crime and violence; e.g. Verneuer, 2020).

This article aims to contribute towards the growing methodical literature on factorial survey designs in a way that makes it easier for researchers without extensive expertise in this area to clearly understand, implement and reflect on design decisions and their consequences for analyses. Because specific (e.g. D-efficient) designs are becoming increasingly popular due to allowing for the practical implementation of vignettes with a large number of dimensions (and/or levels), i.e. a large overall vignette-universe, it is important that clear design categories (for important features such as orthogonality) are offered – to prevent avoidable mistakes during the design-stage.

The findings in this article are new in the sense that they are not intuitively drawn from previous literature, although experts sometimes take them into account automatically. This article exemplifies “new” important aspects that foreground both the “benefits” and “dangers” of D-efficient designs to help researchers to (I.) optimize their designs (e.g. reduce sample sizes, avoid misspecifications) as well as (II.) reflect on their designs and possible implications for their analyses as well as interpretation of results.

In factorial surveys, it is common to divide the vignettes into sets and present each respondent with such a set, thus, gaining more than one response per person for the dependent variable(s). Hence, a smaller number of respondents is required (cf. Atzmüller & Steiner, 2010). In combination with this or on its own, a smaller vignette-sample, a “fraction”, is sometimes drawn from the overall universe. This can be extremely useful for reducing the number of respondents that are required but it also implicates new aspects that must be addressed in depth to ensure that e.g. the internal validity of the experiment is upheld and estimates in the analysis are not biased as a result of the design (Auspurg & Hinz, 2015). To date, random proce-

Direct correspondence to

Julia Kleinewiese, Mannheim Centre for European Social Research (MZES),
University of Mannheim, 68131 Mannheim, Germany.
E-mail: kleinewiese@uni-mannheim.de

dures have been the ‘go to’ method for allocating vignettes to sets and for drawing vignette-samples.

Nevertheless, drawing random vignette-samples comes with some drawbacks, which can be especially meaningful when the sample size is relatively small (see e.g. Auspurg & Hinz, 2015); with random designs, we have no means of controlling two of the important properties of the experimental design: Orthogonality and level balance. To tackle this issue, some researchers have been examining and applying quota designs, including D-efficient designs. Such methods have been widely investigated and applied in a related type of survey experiment, in discrete choice experiments (DCEs), which are applied mostly by economists (e.g. Louviere et al., 2000; see e.g. Gundlach et al. [2018] for a sociological DCE). Unlike the case of DCEs, for factorial surveys, the research on this method of selecting, for example, a sample of vignettes from the universe is very limited (but see e.g. Auspurg & Hinz, 2015; Dülmer, 2015, 2007; Kuhfeld, 2003). Much of this research has focussed on comparing D-efficient designs with random sample designs. Such research is highly valuable for illuminating what the advantages and drawbacks of each procedure are. This article seeks to take this research as a point of departure from which to present accessible methodological information on D-efficient designs.

Currently, there is only a small amount of research that focuses specifically on D-efficiency in factorial surveys. The examination of the method as well as its procedures has focused primarily on comparing different features (including D-efficiency itself) of D-efficient and random vignette-samples (e.g. Auspurg & Hinz, 2015; Dülmer, 2007). The current article builds on such previous research. The two research aims are specifically oriented towards the concept of design resolution: (1) To discuss discrepancies between conceptualizations of resolution IV designs and their implementability with SAS¹; (2) to examine how the resolution – III, IV (with 1, 3 or 5 two-way interactions orthogonalized) and V – affects the minimum size of the vignette-sample that is necessary to still achieve an adequate level of D-efficiency (over 90).

It is for the aforementioned reasons that the current focus is placed on design resolution of D-efficient factorial surveys. The theory section gives a general overview of the research field on factorial survey D-efficiency while the application section exemplifies issues regarding design resolutions of D-efficient vignette-samples and their implications for confounding as well as sample size (e.g. Kuhfeld, 2003). The steps are as follows: First, a brief overview of factorial survey designs is presented, introducing factorial surveys and describing different methods for drawing samples of vignettes as well as the concomitant (dis-)advantages. This is followed,

1 The SAS-version that I am referring to in this article is “SAS OnDemand for Academics – SAS Studio” which is a cloud/online version. Since it is free for all academics it is used very frequently. I refer to it as “SAS” but the assessments and comments made may not hold true for other versions of the software.

by a “state of the art” section on D-efficiency in factorial surveys that includes the theoretical premises that are of importance for the subsequent section: The application (using SAS-macros). This section focuses on the discrepancies between conceptualization and SAS-implementation regarding confounding (aliasing) structures of “resolution IV” designs and on how design resolution impacts the size of the smallest possible vignette-sample that can be constructed from a given full factorial. This section is succeeded by a discussion of the results and a brief conclusion that presents the general implications and recommendations for future research on and with D-efficient factorial survey designs.

Factorial Survey Designs

Factorial Survey Methodology

A factorial survey systematically varies dimensions in scenarios and presents the resulting vignettes to respondents (e.g. Auspurg et al., 2015; Wallander, 2009; Steiner & Atzmüller, 2006; Beck & Opp, 2001; Alexander & Becker, 1978). A parallel between factorial surveys and experiments lies in the condition that the researcher controls the “treatments” (dimensions), so that they can be measured independent of each other (cf. Auspurg et al., 2009; Rossi & Anderson, 1982). By means of varying the dimensions’ levels, the factorial survey allows direct deductions concerning the dependent variable’s variations, as the effects of unobserved variables are eliminated (Dickel & Graeff, 2016). All levels of a dimension need to be clearly distinct from each other. The vignette universe (vignette population/full factorial) is made up of all the vignettes resulting from each possible combination of the dimensions’ levels (Auspurg & Hinz, 2015; Atzmüller & Steiner, 2010; Rossi & Anderson, 1982). In order to avoid dimensions that are composites of a number of attributes, high numbers of dimensions must be selected for some factorial surveys (see Hainmueller et al., 2014). Furthermore, some studies need a high number of levels for one or more dimensions (e.g. due to content-related or analyses requirements). A large number of dimensions and/or levels quickly leads to a very large vignette universe.

The dependent variable is frequently measured on a scale, as a rating score in response to a question (Dickel & Graeff, 2016) regarding the vignette. Frequently, 11-point scales are used (Dülmer, 2014; Wallander, 2009). Usually, additional respondent-specific data are collected and can be included in the analysis of the vignette evaluations (Steiner & Atzmüller, 2006). The aim of statistical analyses is determining the effect of each dimension (and often some interactions) in regard to the respondents’ judgments as well as identifying and explaining the differences

between respondents or groups of respondents (Auspurg & Hinz, 2015; Auspurg et al., 2015; Steiner & Atzmüller, 2006; Beck & Opp, 2001).

If sufficient numbers of respondents are available, every vignette from a vignette universe should be judged by at least five respondents (because if e.g. only one person rates a vignette it is completely confounded with their personal features) (Auspurg & Hinz, 2015). However, with a rising number of vignettes, it becomes increasingly difficult to recruit the necessary number of respondents. There are two solutions which have been prioritized in factorial survey applications, separately or in combination: (1) dividing the overall number of vignettes into sets of equal size or (2) selecting only a sample of the vignettes from the universe (cf. Steiner & Atzmüller, 2006).

Forming sets (decks/blocks) with a specific number of vignettes has the advantage that one can greatly reduce the number of respondents required. With this proceeding, each respondent only answers one set of vignettes. Vignettes can be assigned to sets through experimental variation or random allocation (with or without replacement) (cf. Steiner & Atzmüller, 2006). For optimal distribution, the vignette universe should be a whole multiple of the set size (number of vignettes per set) (Atzmüller & Steiner, 2017; Auspurg & Hinz, 2015). As respondents presumably differ in their assessment tendencies, the measurements are not independent across all vignette-responses. The equivalent variance component is incorporated in the statistical analysis through the modelling of a set effect. Auspurg and Hinz (2015) state: “[...] some parameters become confounded with deck effects [...] but] When all decks are rated by several respondents [...] these parameters remain identifiable in estimations across respondents” (p. 39).

There are several methods for selecting a sample of vignettes from the universe. Steiner and Atzmüller (2006) argue that in the case of randomly drawn (sets of) vignettes, a very complex interaction structure is formed, which may lead to considerable interpretation problems in regard to the estimable effects; they declare that the common implicit assumption that the interaction effects mixed in the effects of interest are equal to zero, is generally an unsatisfactory solution.

This brief introduction to the current state of factorial survey research – particularly its design – provides a basis for understanding the particular methodical design-aspects that are of relevance to the goals of this article. In order to provide insights into other design options and what (not) to do, a number of aspects regarding the design of vignette studies will, subsequently, be described in more detail. All of this shows why the current aims are relevant and provides sufficient knowledge for comprehension of the applied section. The following section gives a brief overview of the different proceedings for drawing a sample of vignettes from the overall vignette universe.

Methods for Drawing Vignette-Samples

There are two important properties of the experimental design regarding the vignette universe as well as vignette-samples: The first property is *orthogonality*. A matrix is orthogonal when the single columns are not correlated with one another. This enables independent (from each other) estimation of the effects of the factors (cf. Auspurg & Hinz, 2015). Thus, for a factorial survey design, orthogonality means that the dimensions (and their interactions) do not correlate with each other (Auspurg & Hinz, 2015; Atzmüller & Steiner, 2010; Taylor, 2006; Rossi & Anderson, 1982); “[...] it enables the researcher to estimate the influence of single dimensions independently of each other” (Auspurg & Hinz, 2015, p. 25). A vignette universe (full factorial) is always orthogonal. The second property is *level balance*. Level balance means that all levels (of every dimension) occur with equal frequencies. Level balance indicates that maximum variance (of the levels) can be used to estimate the effect of each dimension, which leads to the lowest standard errors and, therefore, maximizes the precision of the parameter estimates (cf. Auspurg & Hinz, 2015).

The Cartesian product of dimensions and levels equals the size of the vignette universe. If each respondent judges all vignettes from a universe, the factors are orthogonal to one another in their composition (Dülmer, 2007). In factorial surveys, each person usually only responds to a selected number of vignettes from the universe. This can be achieved through dividing the universe into vignette-sets of equal size (“blocking”) and presenting each respondent with one set only, otherwise, by selecting a sample of vignettes from the universe (or both of the aforementioned).

This section focusses on drawing samples from a vignette universe – presenting methods for drawing such vignette-samples. There are two categories into which techniques for attaining samples fall: (1) *Random samples* are predominantly used to attain a vignette-sample from the universe (the aim is to represent its possible level combinations as closely as can be achieved), however, (2) *quota designs* can also be applied (Dülmer, 2007).

(1) *Random samples* can be drawn once (in sets) and then judged by several respondents (*clustered random design*) or they can be drawn uniquely for each respondent (*simple random design with or without replacement*). The former procedure ensures – given a sufficient number of respondents – several ratings of each included vignette (cf. Jasso, 2006). Each of these strategies has its advantages and its disadvantages. Drawing only once and presenting the resulting sets to several respondents is advisable when one is interested in respondent-specific variation in the vignette-judgements. However, a wider overall portion of the vignette universe is very likely to be achieved when a unique deck is drawn for each respondent (Jasso, 2006).

(2) *Quota samples* are commonly used in conjoint analysis and discrete choice experiments (cf. Dülmer, 2007). There are two types that have been applied frequently: *Fractional factorial designs* (e.g. Marshall & Bradlow, 2002) and *D-efficient designs* (e.g. Kuhfeld et al., 1994). In both variants, the vignette-sample is drawn only once (and then usually divided into sets). Quota sampling utilizes the available knowledge on the statistical properties of the universe in order to select the vignette-sample (of a given size) that most closely/ideally upholds these properties (cf. Dülmer, 2007).

A fractional factorial design is a *symmetrical orthogonal design* when the vignette universe properties of equal level frequencies (symmetrical/balanced) and orthogonality of all factors (e.g. dimensions, interactions) are upheld. It is an *asymmetrical orthogonal design* when it does not have absolute level frequency but preserves orthogonality because one dimension's levels occur with proportional frequency to the other dimensions' levels (Dülmer, 2007).

D-efficient designs relax the rule that a (sample-)design must be perfectly orthogonal. Symmetrical orthogonal designs (perfect level balance as well as orthogonality) represent the vignette universe most closely and minimize parameter estimates' variance. D-efficiency chooses symmetrical orthogonal designs as a point of reference and is, thus, "a standard measure of goodness" (Dülmer, 2007, p. 387) for jointly assessing both orthogonality and level balance, which increases the precision of estimates of the parameters in statistical analyses (Auspurg & Hinz, 2015).

Designs that have a D-efficiency of 100 are also (fractional factorial) symmetrical orthogonal designs (Dülmer, 2007) because they are orthogonal and exhibit level balance. When this is not the case, the best "compromise" between the aims of orthogonality and level balance is searched for (D-efficiency will then be lower than 100). When orthogonal coding has been applied to the vignettes, the range of D-efficiency is 0-100 (see e.g. Dülmer, 2007; Kuhfeld, 1997; Kuhfeld et al., 1994).

There are a number of 'pros and cons' regarding the methods that can be used for selecting a subsample from a vignette universe. From the statistical perspective, reasons why quota designs should be favoured over random designs are their higher efficiency, reliability and power. However, these arguments are primarily applicable for studies that use a fairly low set size, where the selected design is highly D-efficient and quite a high unexplained inter-respondent heterogeneity is to be expected. On the other hand, quota designs can be less valid than random designs; this is most likely when using designs with a low resolution (Dülmer, 2007). In consequence, what type of design is the most expedient for a study can vary – depending on, for instance, the respondent sample and the amount of resources available for implementing the survey. In the past, a majority of factorial survey studies used random designs (Wallander, 2009). However, increasingly D-efficient designs are becoming

more popular. The remaining sections of this article, therefore, focus exclusively on D-efficient designs.

D-Efficiency

Taking the preceding overview as a point of departure, this section provides a more in-depth elaboration of D-efficiency. It begins with a general section on D-efficient designs that is followed by subsections on sample size as well as design resolution. This constitutes the final theoretical building block for assessing the implications in the applied section.

D-Efficient Designs

When one applies (D-)efficiency-maximizing methods for finding a suitable vignette-sample, one should be able to reach the same amount of precision as with random sampling but with fewer respondents and/or vignettes per set. Moreover, it can be easier to reach and assess the goal that all parameters of interest can be identified. Against this background, an objective is to find a fraction of the vignette universe with maximal gain of information, about all parameters that are of relevance for the research aim(s).

The previously described combination of considering both orthogonality and level balance can be specified in regard to optimizing D-efficiency: The goal is maximizing the variance of the dimensions' levels while simultaneously minimizing the correlations between the factors (e.g. dimensions, interactions). The equivalent optimums are level balance and orthogonality.

D-efficiency contains (is reliant on) the Fisher Information Matrix (FIM) [$X'X$], where X indicates a vector (of vignette variables) (Auspurg & Hinz, 2015; Kuhfeld et al., 1994). There are other measures of efficiency (such as A-efficiency; a function of the arithmetic mean of the $X'X$ matrix) than D-efficiency, which is based on the geometric mean of the matrix. However, these efficiency measures are usually highly correlated with each other and D-efficiency is used most frequently (Auspurg & Hinz, 2015; Kuhfeld, 1997). The formula for D-efficiency is as follows [p =parameters to be estimated (including the intercept); n_s =number of vignettes in the fraction; $|X'X|=FIM$]:

$$\text{D-efficiency} = 100 \times \frac{1}{n_s \times |(X'X)^{-1}|^{\frac{1}{p}}} = 100 \times \left(\frac{1}{n_s} \times |X'X|^{\frac{1}{p}} \right)$$

(Auspurg & Hinz, 2015; Dülmer, 2007; Kuhfeld et al., 1994)

Fewer dimensions (or other estimated parameters e.g. 2-way-interactions) reduce the correlation of parameters with each other. Larger vignette-samples, from a vignette universe of a given size, (sample-sizes prescribe the degrees of freedom for parameter estimates) decrease covariation (and, therefore, correlations) between dimensions, increasing precision of parameter estimates. The FIM reflects how high the information is for parameter estimates. The information matrix is the inverse of the variance-covariance matrix.

As stated in subsection *Methods for drawing vignette-samples*, when the dimensions' levels from a universe are in orthogonal coding, the maximum D-efficiency that can be reached is 100. To elaborate upon this: Methodical literature states that a D-efficiency over 90 should be sufficient for experimental survey designs in the social sciences (Auspurg & Hinz, 2015).

The more efficient a design is, the fewer vignette-judgements one requires to achieve the same (level of) statistical power:

Efficiencies are typically stated in relative terms, as in design A is 80% as efficient as design B. In practical terms this means you will need 25% more (the reciprocal of 80%) design A observations (respondents, choice sets per respondent or a combination of both) to get the same standard errors and significances as with the more efficient design B. (*Chrzan & Orme, 2000, p. 169*)

Sample Size

When the size of a vignette-sample from a given universe increases it becomes more likely that one can reach a high D-efficiency. Auspurg and Hinz (2015) state that this leads to a trade-off because – given a fixed number of respondents and set size – the number of respondents per set decreases. However, an additional option is that one could increase the set size (even if this can increase the design effect; for more information on the design effect see e.g. Auspurg & Hinz, 2015, pp. 50-55).

The smallest possible vignette-sample is the number of parameters that are to be estimated plus one. The smallest sample is normally very inefficient and does not fulfill the criteria of a D-efficiency over 90 (cf. Auspurg & Hinz, 2015).

In Auspurg and Hinz' (2015) comparison of random vignette-samples and D-efficient vignette-samples, using two different vignette universes, the D-efficient designs are always more efficient. The differences are especially high for small vignette-samples and decrease as the sample size increases. The maximum correlations of the random samples are much higher than those of the D-efficient samples, meaning that the random samples' dimensions (experimental factors) lose much of their independency, threatening internal validity. D-efficient samples usually exhibit higher variance (of levels within each dimension), which means higher sta-

tistical power for correctly identifying the effects of the dimensions. Due to hardly any randomness in the selection of the vignette-sample, the variation in the D-efficiency of (same-sized) D-efficient vignette-samples over several “tries” is very low in comparison to the variation exhibited by random samples.

Design Resolution

While small vignette sample size with a D-efficiency of 100 ensures that the dimensions (main effects) are orthogonal to each other and have level balance (in estimation: standard errors are minimized; statistical efficiency is maximized), this is still likely to lead to biased estimates if relevant two-way (or higher) interactions are not negligible. If such interaction effects are not specified in a design, but do have an effect, this leads to confounding of main and interaction effects. This can bias the estimations of the main effects and rules out the estimation of the interaction effects. If main effects are biased, this leads to biased (in some cases entirely false) interpretations of the data (Auspurg, 2018). For this reason, it is important to consider, which effects have been orthogonalized in a D-efficient design. Commonly, this has been approximated by applying the categorization of designs into “resolutions”.

Resolution identifies which effects are estimable. For resolution III designs, all main effects are estimable free of each other, but some of them are confounded with two-factor interactions. For resolution IV designs, all main effects are estimable free of each other and free of all two-factor interactions, but some two-factor interactions are confounded with other two-factor interactions. For resolution V designs, all main effects and two-factor interactions are estimable free of each other. (Kuhfeld, 2003, p. 237)

D-efficiency is always measured relatively to the selected design resolution. When the orthogonally coded levels of a dimension (in a given vignette-fraction) are completely identical with those of a 2-/3-/4-way interaction then, statistically, they are entirely correlated and their effects cannot be separated in the analysis i.e. they are completely “confounded” or “aliased”. This can also include the intercept. The coefficients of main effects that are aliased with interaction effects are only estimable (unbiased) if those interaction effects have no effect (effect = 0) on the dependent variable. If the wrong assumptions are made this results in biased estimates of the (main) effects (cf. Auspurg & Hinz, 2015).

In marketing research, resolution III designs (also termed “orthogonal arrays”) are mostly used (Kuhfeld, 2003). However, in the social sciences, one should always consider possible two-way interactions that might have an effect (e.g. Auspurg, 2018). While, therefore, it seems advisable to use resolution V designs in sociological research, to date, resolution IV designs have usually been applied. This

can be sufficient, however, when using resolution IV designs the researcher should be aware that this might cause biased results if they err in the assumption that the confounded interactions are negligible.

The rules for which level of factors can be estimated independently from one another (or part of the factors from a level; e.g. resolution IV designs) differ, depending on whether or not the number of the resolution (r) is (1) odd – e.g. resolutions III and V – or (2) even – e.g. resolution IV. The general rule for the former (1) is that all effects of order $e = (r - 1)/2$ or below are estimable independently from one another but at least some of the effects of order e are aliased with interactions of order $e + 1$. In the latter case (2) the rule is slightly different: Effects of order $e = (r - 2)/2$ are estimable independently from one another and also from interactions of order $e + 1$ (Kuhfeld, 2005).

Much previous research has been conducted under the assumption that higher resolutions will always necessitate designs with larger vignette-samples (e.g. Kuhfeld, 2003, 2005). Moreover, research has increasingly questioned the primacy given to maximizing the efficiency of designs, arguing that unbiased estimation of effects should be the superior goal (Auspurg, 2018; Czymara & Schmidt-Catran, 2018). Minimizing (possible) bias in estimation of effects requires using higher design resolutions.

Application with SAS-Macros

This section turns towards a practical application, examining examples of implementations of D-efficient factorial survey designs using SAS-macros. D-efficient factorial survey designs in the social sciences are normally constructed by means of computer algorithms. In sociology, the SAS-macros written by Warren F. Kuhfeld (for more details see e.g. Kuhfeld, 2003) are commonly used (see e.g. Auspurg & Hinz, 2015; Dülmer, 2007). These macros enable the computation of D-efficient samples and sets as well as pre-construction assessment and post-construction evaluation. A number of details regarding the design can also be evaluated in varying detail (e.g. correlations, aliasing structure).

Proceedings: The Design and the Macros

I used the SAS-macros `%mktruns` and `%mktex` to test my propositions. My first aim was to use the SAS macros to try and construct resolution IV designs that fulfil the conceptual requirements that Kuhfeld (2003, p. 237) defined. I selected a $2^8 = 256$ vignettes universe because I presume that this simple structure is very useful for assessing aliasing structures. I used the macros to construct a fraction with a D-efficiency of 100, thus, 0 violations (of orthogonality and level balance)

and a sample size of $n=16$ vignettes. I included one two-way interaction-effect to be orthogonalized (x_1*x_2). I documented the aliasing scheme of the design, in order to be able to assess whether or not the properties postulated in literature are present. I then repeated this procedure for two more resolution IV designs – one design with 3 two-way interactions (x_1*x_2 x_2*x_3 x_3*x_4) and one with 5 two-way interactions (x_1*x_2 x_2*x_3 x_3*x_4 x_4*x_5 x_5*x_6). Both of these designs also had a D-efficiency of 100, 0 violations and a vignette-sample size of $n=16$.

For my second research aim, which focusses on the relationship of design resolution and sample size, I selected three vignette universes: $4^4 = 256$, $4^4 2^1 = 512$ and $4^4 2^2 = 1024$. The number of dimensions and their levels for the first universe were selected due to a specific research interest in this structure and the others each add one more two-level dimension, causing each universe to be twice as large as the previous one. This, of course, could have been done differently. I searched for the smallest possible vignette-sample with a D-efficiency as close as possible to 100 for each universe. I constructed five designs for each universe, documenting the sample size, the D-efficiency and the number of violations. The first fraction for each universe was a resolution III design. The second, third and fourth designs were always of resolution IV (according to SAS). Three designs were selected from resolution IV because this resolution can be used to describe the inclusion of various numbers of two-way interactions, from merely one (more than resolution III) to all but one (less than resolution V). The objective was to see if there is a large difference between the minimum sample sizes of resolution IV designs, depending on how many interactions are fixed as orthogonalized in a design. The first of these designs orthogonalizes one two-way interaction (x_1*x_2), the next design three two-way interactions (x_1*x_2 x_1*x_3 x_1*x_4) followed by a design with five two-way interactions (x_1*x_2 x_1*x_3 x_1*x_4 x_2*x_3 x_2*x_4). These three steps were chosen because five is the highest number of two-way interactions possible in the first vignette universe as a resolution IV design (since that is one below 6 two-way interactions, which would be a resolution V design for the first vignette universe). As a final step, a resolution V design was computed and documented for each vignette universe.

Results

Resolution IV Aliasing

Regarding the three resolution IV designs that were supposed to be computed for the first research aim, I find that no main effects are aliased with one another. Furthermore, orthogonalized interactions are not aliased with main effects or other orthogonalized interactions. However, some other two-way interactions are aliased with main effects, orthogonalized interactions and interactions that were not specified to be orthogonal. This, in effect, does not qualify the designs to be of resolu-

tion IV (for this, none of the two-way interactions should be aliased with any main effects) but rather only to be of resolution III. The result is that for this vignette universe, it would have only been possible to use SAS to compute a resolution III or a resolution V design, in accordance with Kuhfeld's definition (2003, p. 237).

Resolutions and Sample Sizes

Table 1 depicts the smallest sample sizes, the D-efficiencies of the samples as well as the violations for each universe when a resolution III design is chosen. For each of the three vignette universes, the sample size is $n=16$ for the smallest possible size with an adequate D-efficiency (over 90 and as close as possible to 100). The fractions in Table 1 all have a D-efficiency of 100 and, therefore, have 0 violations (of orthogonality and level balance).

Table 2 gives an overview of the sample sizes, the D-efficiencies of the vignette-samples and the violations for the three designs that fall into the category "resolution IV". As shown below, the first and second design-types (1 and 3 two-way interactions) are the same across all vignette universes and in regard to each other. The smallest possible sample size always consists of 64 vignettes, has a D-efficiency of 100 with 0 violations. The final resolution IV design-type (with 5 two-way interactions) has a larger number of vignettes for the smallest sample sizes possible than the first two types. However, it remains the same across the vignette universes. The size is $n=128$ with a D-efficiency of 96 (with a slight variation in the second decimal place) and 1 violation for each vignette universe.

Table 3 presents the smallest possible vignette-sample sizes, with their D-efficiencies and violations for resolution V designs of each of the three vignette universes. The sample sizes of the first two universes are $n=128$ with a D-efficiency of 95 (with some variation in the first and second decimal places) and 1 violation. For the last and largest universe, it was not possible to compute a sample of that size with a D-efficiency over 90. The smallest sample size that is possible and fulfils this criterion is $n=256$. It has a D-efficiency of 100 with 0 violations.

Table 1 Smallest vignette-sample sizes with resolution III design

Universe (from dimensions & levels	Number of dimensions	No. of 2-way interaction-factors	Resolution III	
			Sample size (n)	Violations
$4^4 = 256$	4	6	16	0
$4^4 2^1 = 512$	5	10	16	0
$4^4 2^2 = 1024$	6	15	16	0

Table 2 Smallest vignette-sample sizes with resolution IV (1/3/5 two-way interactions) designs

Universe (from dimensions & levels	Number of dimensions	No. of 2-way interaction-factors	Resolution IV											
			1 two-way interaction-factor			3 two-way interaction-factors			5 two-way interaction-factors					
			Sample size (n)	D- efficiency	Violations	Sample size (n)	D- efficiency	Violations	Sample size (n)	D- efficiency	Violations	Sample size (n)	D- efficiency	Violations
$4^4 = 256$	4	6	64	100	0	64	100	0	128	96.12	0	128	96.12	1
$4^4 2^1 = 512$	5	10	64	100	0	64	100	0	128	96.13	0	128	96.13	1
$4^4 2^2 = 1024$	6	15	64	100	0	64	100	0	128	96.16	0	128	96.16	1

Table 3 Smallest vignette-sample sizes with resolution V design

Universe (from dimensions & levels	Number of dimensions	No. of 2-way interaction-factors	Resolution V		
			Sample size (n)	D- efficiency	Violations
$4^4 = 256$	4	6	128	95.21	1
$4^4 2^1 = 512$	5	10	128	95.68	1
$4^4 2^2 = 1024$	6	15	256	100	0

Discussion of the Results

It is commonly assumed that when designs are of resolution IV, all main effects are estimable independently from one another and from all of the two-way interactions, while two-way interactions may be aliased with each other (e.g. Kuhfeld, 2003). My results offer new insights into the computational issues (SAS) of constructing resolution IV designs. The computed designs concur with the definition in previous literature in that no main effects are aliased with one another and that the orthogonalized two-way interactions are not aliased with the main effects. Also, some two-way interactions that are not orthogonalized are confounded with other two-way interactions. However, a discrepancy between conceptualization and implementation arises: Some non-orthogonalized two-way interactions are aliased with main effects, which may cause estimates of the main effects to be biased. This means that looking at the aliasing structures of the aspiring resolution IV designs shows that they do not fulfil all theoretical requirements and must, instead, be defined as resolution III designs. This suggests that the “catch all” category (resolution IV) between the clearly defined resolutions III and V needs to be treated with caution in implementation. If possible, I suggest that researchers select a resolution V design. If not, aliasing schemes must be carefully monitored (and reported as supplementary material to publications – for reasons of transparency).

Regarding the results on how resolutions impact the smallest possible vignette samples with an adequate D-efficiency, first some general observations: Within each resolution (or subcategory in the case of resolution IV) the smallest sample size is the same for all vignette universes (except for the largest universe in resolution V); even though universes two and three each have one dimension more and are twice as large as the directly preceding (smaller) universe. This is an interesting finding because the same sample size is relatively a smaller fraction when the universe is larger, for example, $n=16$ is (relatively) a smaller fraction of the vignette universe for design three ($1/64$) than for the second largest universe ($1/32$) and the smallest universe ($1/16$). It is also noteworthy that all sample sizes are whole multiples of each other, of the dimensions as well as their levels and that the universes are whole multiples of the samples. This is due to the structure of the full factorial and may not be so clear cut in the case of, for example, samples of vignettes from universes that are made up of dimensions whose numbers of levels are not multiples of one another. Violations are always equal to 0 when D-efficiency is at 100. This is because that amount of D-efficiency requires perfect orthogonality and level balance. There is a noticeable difference in sample sizes across the resolutions (and subcategories). Resolution III has a 16-vignette D-efficient sample. If one interaction is included (resolution IV, category 1) then the minimum-sample is four times larger ($n=64$) than in the resolution III designs. For resolution IV with 3 two-way interactions, category 2, the sample size remains at $n=64$. However, for resolution IV, category 3, with 5 interactions, the sample size ($n=128$) is twice as

large as for the first two categories in the resolution and eight times as large as in the resolution III designs. Comparing designs of the three categories of resolution IV fractions one can, therefore, claim that there are substantial differences between some (but not all) differing numbers of orthogonalized interactions in regard to the minimum sample size within the resolution. For the third universe with resolution V there is no subsample of vignettes that is smaller than 256 and has a D-efficiency of over 90. A difference in the sample size between the universes is present only for this resolution. The results suggest that a larger vignette universe does not have to increase the smallest possible sample size. Moreover, a higher resolution does not have to increase the smallest possible sample size. Interestingly, the resolutions do not necessarily determine the boundaries at which the minimum sample sizes increase.

Conclusions

The current application provides an added value for vignette-design methodology: As a first step, it examines structural properties of computed (SAS-macros) designs and, for two important issues pertaining design resolutions, compares the results shown in the computed designs with the assertions from the literature. The conclusions provide a basis for future computational (SAS-macros) or mathematically-driven research on design resolution and sample sizes of D-efficient designs. Although the results can lead only to tentative conclusions they should lead to further extensive exploration of this topic.

Some central deductions drawn from the conducted research are: (1) The examined aliasing structures indicate a discrepancy between previous definitions and the aliasing structures of designs resulting from SAS-macros. (2) For the selection of a small sample size, the overall size of the vignette universe does not necessarily play a fundamental role, rather the dimensions' level-combinations. It should be considered from the early stages of design onwards that smallest sample sizes with an adequate D-efficiency can vary strongly depending on the combinations of numbers of dimension-levels that are chosen. When all dimensions have the same number of levels or the level-number of a part of the dimensions is a whole multiple of the other dimensions' level-number smaller vignette-samples can reach an adequate D-efficiency than with more irregular combinations. (3) There is a trade-off between a minimal vignette-sample size and number of orthogonalized factors (not necessarily resolutions). (4) Resolution V designs with an implementable sample size are often possible. Therefore, it is highly recommendable to apply resolution V designs. Sometimes, however, this may not be implementable in research practice, leading researchers to apply resolution IV designs. When implementing resolution IV designs, one should always state precisely, which interactions have been orthog-

onalized. Furthermore, especially when using computer algorithms (e.g. SAS-macros), one must assess the aliasing structures of the design in order to determine if the output design fulfils all of the theoretically presumed orthogonalizations.

Of course, these suggestions are more implementable for some vignette studies than others. Studies, for example, on situational, deviant actions (e.g. Kleinewiese & Graeff, 2020; Wikström et al., 2012) often have more flexibility when it comes to selecting the exact numbers of dimensions' levels. Studies on other topics, such as the gender-pay-gap (e.g. Auspurg, Hinz & Sauer, 2017), may include dimensions (e.g. gender) in which the number of levels is not so easily alterable.

Put in a nutshell, this article clearly shows that D-efficient designs are suitable and expedient for a majority of factorial survey studies – even for researchers without prior “expert knowledge” on experimental survey methodology. It exemplifies, how small changes in design can have large implementation-advantages regarding sample sizes and aliasing. At its core, it reflects upon previous common usage of “resolution IV designs”, showing the potential drawbacks of this approach. Based on the conceptual and applied sections, it advises making the usage of resolution V designs a standard in social science research. It supports the necessity of improving transparency regarding research designs. This is important because researchers, reviewers, publishers and readers should have a clear comprehension of the design and its implications for the analyses and the interpretation of the results.

Taking this as a point of departure, future studies should systematically examine the proposed examples (e.g. via comparisons with random samples) to provide further support for the suggested proceedings. Another interesting design-aspect requiring further examination is the interrelation of vignette sampling and blocking. While previous research shows that D-efficient blocking of vignettes to sets leads to less biases in effect estimates than random blocking (Su & Steiner, 2020), as a next step, it would be important to further examine the interrelations of sampling and blocking (both D-efficient and random), especially regarding implementation and possible issues.

References

- Alexander, C. S., & Becker, H. J. (1978). The use of vignettes in survey research. *Public Opinion Quarterly*, 42(1), 93-104.
- Atzmüller, C., & Steiner, P. M. (2010). Experimental vignette studies in survey research. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 6(3), 128-138.
- Atzmüller, C., & Steiner P. M. (2017). Was ist ein faktorieller Survey?. In M. W. Schnell, C. Schulz, C. Atzmüller, & C. Dunger (Eds.), *Ärztliche Wertehaltungen gegenüber nichteinwilligungsfähigen Patienten* (pp.29-52). Wiesbaden: Springer.

- Auspurg, K. (2018). Konfundierte Ergebnisse durch ein zu stark beschränktes Design?. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 70(1), 87-92.
- Auspurg, K., Hinz, T., & Liebig, S. (2009). Komplexität von Vignetten, Lerneffekte und Plausibilität im Faktoriellen Survey. *Methoden – Daten – Analysen*, 3(1), 59-96.
- Auspurg, K., & Hinz, T. (2014). *Factorial survey experiments* (Vol. 175). Sage Publications.
- Auspurg, K., Hinz, T., Liebig, S., & Sauer, C. (2015). The factorial survey as a method for measuring sensitive issues. In U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, & P. Sturgis (Eds.), *Improving survey methods: Lessons from recent research* (pp. 137-149). New York: Routledge.
- Auspurg, K., Hinz, T., & Sauer, C. (2017). Why should women get less? Evidence on the gender pay gap from multifactorial survey experiments. *American Sociological Review*, 82(1), 179-210.
- Beck, M., & Opp, K. (2001). Der faktorielle Survey und die Messung von Normen. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 53(2), 283-306.
- Chrzan, K., & Orme, B. (2000). An overview and comparison of design strategies for choice-based conjoint analysis. *Sawtooth software research paper series*, 98382.
- Czymara, C. S., & Schmidt-Catran, A. W. (2018). Konfundierungen in Vignettenanalysen mit einzelnen defizienten Vignettenstichproben. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 70(1), 93-103.
- Dickel, P., & Graeff, P. (2016). Applying factorial surveys for analyzing complex, morally challenging and sensitive topics in entrepreneurship research: The case of entrepreneurial ethics. In E.S.C Berger, & A. Kuckertz (Eds.), *Complexity in entrepreneurship, innovation and technology research* (pp. 199-217). Cham: Springer.
- Dülmer, H. (2007). Experimental plans in factorial surveys: random or quota design?. *Sociological Methods & Research*, 35(3), 382-409.
- Dülmer, H. (2014). Vignetten. In N. Baur, & J. Blasius (Eds.), *Handbuch Methoden der empirischen Sozialforschung* (pp. 721-732). Springer VS, Wiesbaden.
- Dülmer, H. (2015). The factorial survey: Design selection and its impact on reliability and internal validity. *Sociological Methods & Research*, 45(2), 304-347.
- Eifler, S. (2007). Evaluating the validity of self-reported deviant behavior using vignette analyses. *Quality & Quantity*, 41(2), 303-318.
- Eifler, S., Pollich, D., & Reinecke, J. (2014). Die Identifikation von sozialer Erwünschtheit bei der Anwendung von Vignetten mit Mischverteilungsmodellen. In S. Eifler, & D. Pollich (Eds.), *Empirische Forschung über Kriminalität* (pp. 217-247). Springer VS, Wiesbaden.
- Engel, U., & Schmidt, B. O. (2019). Unit- und Item-Nonresponse. In N. Baur, & J. Blasius (Eds.), *Handbuch Methoden der empirischen Sozialforschung* (pp. 385-404). Springer VS, Wiesbaden.
- Gundlach, A., Ehrlinspiel, M., Kirsch, S., Koschker, A., & Sagebiel, J. (2018). Investigating people's preferences for car-free city centers: A discrete choice experiment. *Transportation Research Part D: Transport and Environment*, 63, 677-688. <https://doi.org/10.1016/j.trd.2018.07.004>
- Hainmueller, J., Hopkins, D. J., & Yamamoto, T. (2014). Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Political analysis*, 22(1), 1-30.

- Hainmueller, J., Hangartner, D., & Yamamoto, T. (2015). Validating vignette and conjoint survey experiments against real-world behavior. *Proceedings of the National Academy of Sciences*, 112(8), 2395-2400.
- Jasso, G. (2006). Factorial survey methods for studying beliefs and judgments. *Sociological Methods & Research*, 34(3), 334-423.
- Kleinewiese, J., & Graeff, P. (2020). Ethical decisions between the conflicting priorities of legality and group loyalty: Scrutinizing the “code of silence” among volunteer firefighters with a vignette-based factorial survey. *Deviant Behavior*, 4(6), 1–14. <https://doi.org/10.1080/01639625.2020.1738640>
- Kuhfeld, W. F., Tobias, R. D., & Garratt, M. (1994). Efficient experimental design with marketing research applications. *Journal of Marketing Research*, 31(4), 545-557.
- Kuhfeld, W. F. (1997). Efficient experimental designs using computerized searches. *Research Paper Series*, SAS Institute, Inc. Cary, NC: SAS Institute Inc.
- Kuhfeld, W. F. (2003). *Marketing research methods in SAS: Experimental design, choice, conjoint, and graphical techniques*. Cary, NC: SAS Institute Inc.
- Kuhfeld, W. F. (2005). Experimental design, efficiency, coding, and choice designs. *Marketing research methods in SAS: Experimental design, choice, conjoint, and graphical techniques*, 47-97.
- Louviere, J. J., Hensher, D. A., & Swait, J. D. (2000). *Stated choice methods: analysis and applications*. Cambridge: Cambridge university press.
- Rossi, P. H., & Anderson, A. B. (1982). The factorial survey approach: An introduction. In P. H. Rossi, & S. L. Nock (Eds.), *Measuring social judgments: The factorial survey approach* (pp. 15-67). Beverly Hills, CA: Sage.
- Steiner, P. M., & Atzmüller, C. (2006). Experimentelle Vignettendesigns in Faktoriellen Surveys. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 58(1), 117-146.
- Su, D., & Steiner, P. M. (2020). An evaluation of experimental designs for constructing vignette sets in factorial surveys. *Sociological Methods & Research*, 49(2), 455-497
- Taylor, B. J. (2006). Factorial surveys: Using vignettes to study professional judgement. *British Journal of Social Work*, 36(7), 1187-1207.
- Verneuer, L. M. (2020). Selbstbericht und Vignette als Instrumente zur empirischen Abbildung von Gewalt als Sanktionshandlung. In I. Krumpal, & R. Berger (Eds.), *Devianz und Subkulturen* (pp. 241-277). Springer VS, Wiesbaden.
- Wallander, L. (2009). 25 years of factorial surveys in sociology: A review. *Social Science Research*, 38(3), 505-520.
- Wikström, P.-O. H., Oberwittler, D., Treiber, K., & Hardie, B. (2012). *Breaking rules: The social and situational dynamics of young people's urban crime*. Oxford University Press.

Factorial Surveys with Multiple Ratings per Vignette. A Seemingly Unrelated Multilevel Regressions Framework

Alexander W. Schmidt-Catran

Institute of Sociology, Goethe-University Frankfurt

Abstract

Factorial surveys are a prominent tool in the social sciences. Reanalyzing a literature survey on the factorial survey approach (Wallander, 2009), I show that about a quarter of applied factorial surveys asks respondents to provide multiple ratings on the same vignette. This paper is the first to propose a statistical modeling approach for precisely this situation. Data from factorial surveys with multiple ratings per vignette are afflicted with two sources of statistical dependencies. First, each respondent answers multiple vignettes, which is typically accounted for via random effects models, and, second, each vignette prompts multiple ratings. The first problem is common for almost any factorial survey and has been addressed decades ago. The second problem is addressed here. I propose to apply a seemingly unrelated regression approach to account for the statistical dependencies between multiple ratings per vignette. Due to the use of a structural equation modeling approach, the model allows not only to correctly compare coefficients across ratings but also to analyze the factor structure underlying these ratings. The proposed model is illustrated by two examples from recent research. All data and syntax are available online and allows for an easy adaption of the proposed model to readers' own research.

Keywords: factorial survey, vignette study, seemingly unrelated regressions, multiple ratings, multilevel, random effects, factor analysis, latent variables



Factorial surveys, also called vignette studies, present artificial descriptions of people, objects or situations (vignettes), which are judged (rated) by survey respondents. Each vignette contains multiple theoretically relevant factors (dimensions) simultaneously. Thereby, factorial surveys allow investigating how the multi-dimensional characteristic of an object, person or situation, affects respondents' attitudes towards it (Jasso, 2019). The characteristics (levels) of the factors are varied systematically across the entire universe of vignettes. Factorial surveys thereby combine the virtues of experimental approaches to causal inference with classical survey research (Atzmüller & Steiner, 2010). They are particularly useful if the characteristics of interest are strongly confounded in reality, or at least in the perception of the respondents, and if the object of interest is suspect to social desirability (Atzmüller & Steiner, 2010; Wallander, 2009).

For example, Czymara and Schmidt-Catran (2016) ask “who is welcome in Germany?” and present descriptions of immigrants to their respondents. The immigrants are described in terms of their education, gender, country of origin, language skills, motivation, and religion. Each of these factors is constituted by multiple levels, for example, immigrants have no religion, are Christian or Muslims. The design allows investigating the relative impact of each dimension on the acceptance of immigrants and the estimation of the effect of specific levels. What is more important, economically relevant characteristics like education, or cultural features like religion? Are Christians preferred over Muslims?

Going back to the seminal work by Rossi and colleagues (Rossi & Nock, 1982), factorial surveys have now been around for 40 years and are frequently used in social science research (for an overview see Wallander, 2009). Many papers have been written about issues of designing and analyzing factorial surveys (for an overview see Jasso, 2006). Methodological issues concern for example the design of the vignettes (Auspurg, Hinz, Sauer, & Liebig, 2015), the assignment of vignettes to respondents (for example Atzmüller & Steiner, 2010; Dülmer, 2007, 2016) or the statistical method for the efficient estimation of the effects of vignette characteristics (for example Hainmueller, Hopkins, & Yamamoto, 2014; Jasso, 2006). In some way, most previous methodological papers focus on how to best deal with the multi-dimensionality of vignette characteristics. This paper takes a different route; it brings the multi-dimensionality of attitudes towards a social object into play.

Typically, factorial surveys require respondents to provide *one* rating per vignette, thereby restricting the measurement of the attitude towards the described

Direct correspondence to

Alexander Schmidt-Catran, Professur für Soziologie mit dem Schwerpunkt Methoden der quantitativen empirischen Sozialforschung, Goethe-Universität Frankfurt, Fachbereich Gesellschaftswissenschaften, Institut für Soziologie, Theodor-W-Adorno-Platz 6, PEG – Gebäude, 60323 Frankfurt
E-mail: alex@alexanderwschmidt.de

object to one dimension. However, factorial surveys with multiple rating questions per vignette, are not uncommon. A re-analysis of the studies discussed in Wallander's (2009) review indicates that about one quarter (27%) of applied factorial surveys measure *multiple ratings per vignette*.¹ More recent examples of such surveys are Harell et al. (2012), Weinberg et al. (2014), Czymara and Schmidt-Catran (2016) and Diehl et al. (2018); the last two of which are used as examples in this article. To the best of my knowledge, no special modeling approach for factorial surveys with *multiple ratings per vignette* has been introduced previously. It is important to define the term "multiple ratings per vignette" in order to avoid misunderstandings. A factorial survey typically provides *multiple vignettes* to a respondent, i.e. multiple descriptions of objects that vary in their characteristics. Thus, we get multiple ratings per respondent—as many as the respondent received vignettes. As discussed below, the hierarchical structure resulting from this survey design, is typically accounted for via multilevel models. Additionally, in some factorial surveys, respondents must provide multiple ratings on each vignette description. This results in *multiple ratings per vignette*. For example, Czymara and Schmidt-Catran (2016) provide 14 descriptions of immigrants to their respondents. On each of these 14 vignettes, respondents had to provide three ratings, resulting in a total of 42 (= 14 x 3) ratings per respondent.

The following paper proposes a statistical model for the analysis of such data. This model can be applied to any data from factorial surveys that (1.) include multiple ratings per vignette (at least 2) and (2.) multiple vignettes per respondent (at least 2). More precisely, the technique proposed here, models each of the ratings as a separate dependent variable and thereby allows for the analysis of their differences and commonalities regarding their determinants.² The basic idea is to use a seemingly unrelated regression framework combined with a structural equation approach to multilevel modeling. This allows for a simultaneous modelling of multiple dependent variables (i.e. the multiple ratings per vignette). Multilevel modeling has been recommended for the analysis of factorial surveys as it accounts for the statistical dependencies in the data, due to the fact that each respondent is confronted with *multiple vignettes* (Hox, Kreft, & Hermkens, 1991). Combining multilevel analysis with the seemingly unrelated regression approach allows correct accounting for the additional statistical dependencies due to the measurement

-
- 1 I want to give special thanks to Lisa Wallander for providing me with the data she collected for her literature survey. My re-analysis of the studies reviewed by Wallander (2009) can be found in the online appendix (Table OA1) of this paper: <http://www.schmidt-catran.de/sumreg.html>.
 - 2 It may be that the multiple ratings per vignette constitute indicators of the same (unidimensional) latent construct. In this case, the multiple ratings may be combined into a single dependent variable before the analysis, rather than using the model proposed here, which is suitable only if the analysis has multiple dependent variables.

of *multiple ratings per vignette*. Finally, the use of structural equation modeling gives the opportunity to analyze the latent structure underlying the ratings.

A Seemingly Unrelated Multilevel Regressions Framework

In a seminal paper, Zellner (1962) proposed a method to estimate seemingly unrelated regression (SUR) models. He discusses how to account for the fact that estimation results from a set of regressions which use different dependent variables but share (some) predictors are statistically not independent. If all regressions have exactly the same set of predictors, this does not affect the estimated model parameters but the statistical tests necessary for comparing parameters across the regressions (Zellner, 1962: 351, 355). If the regressions differ not only in their dependent but also in their independent variables, accounting for the statistical dependence does also directly affect the estimators (Zellner, 1962: 351).

In the context of factorial surveys with multiple ratings per vignette, each dependent variable (i.e. rating) will always be dependent on the same vignette dimensions (i.e. predictors) by design. Hence, when analyzing the impact of the vignette dimensions only, estimating seemingly unrelated multilevel regressions (SUMREG) provides the same estimators as separate regressions. In that case, the SUR approach boils down to a multivariate regression model, which can be seen as a special case of the former. Nevertheless, accounting for the statistical dependence of the estimators, more precisely of the error terms, is important when comparing coefficients across dependent variables (Zellner, 1962: 355).

If respondent-level characteristics are added to the set of predictors, it may be that there are theoretical reasons to include some variables in one equation but not in another (see Example 1.3 below). In that case, the SUR approach will yield different (more efficient) estimates than a separate regression approach. Nevertheless, the emphasis in this article is on the more likely case of identical predictors in all regressions and therefore on comparing coefficients across them.

In his seminal paper, Zellner (1962) proposed a two-stage approach to the “efficient” estimation of SURs. However, the model can also be estimated in one step, using structural equation modeling. Formally the model can be understood as a system of i regression equations:³

3 Note that the index i here indicates regression equations—not units of analysis—because the regression equations are presented in matrix notation, which does not include an index for the units of analysis.

$$\begin{aligned}
 Y_1 &= X\beta + e_1 \\
 &\vdots \\
 Y_i &= X\beta + e_i,
 \end{aligned}$$

in which the error terms are allowed to be correlated across equations. Thus, the variance-covariance matrix of the error terms is an unrestricted matrix in which the error variances are located at the diagonal and their covariances at the off-diagonals:

$$\Sigma_e = \begin{bmatrix} \sigma_{11} & \dots & \sigma_{1i} \\ \vdots & \ddots & \vdots \\ \sigma_{i1} & \dots & \sigma_{ii} \end{bmatrix}.$$

This model can be extended to account for multiple error components (Baltagi, 1980), i.e. a multilevel structure in the data:

$$\begin{aligned}
 y_1 &= X\beta + u_1 + e_1 \\
 &\vdots \\
 y_i &= X\beta + u_i + e_i,
 \end{aligned}$$

where each component of the composite errors $\varepsilon_i (= u_i + e_i)^4$ allows correlations across equations but the components are independent of each other:

$$\Sigma_\varepsilon = \begin{bmatrix} \sigma_{11}^u & \dots & \sigma_{1i}^u & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{i1}^u & \dots & \sigma_{ii}^u & 0 & \dots & 0 \\ 0 & \dots & 0 & \sigma_{11}^e & \dots & \sigma_{1i}^e \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \sigma_{i1}^e & \dots & \sigma_{ii}^e \end{bmatrix}$$

Such a model, traditionally employed for the analysis of panel data, seems to be perfectly suited to analyze factorial surveys with multiple ratings per vignette.⁵ One argument for this has been laid out above. The model accounts for the statistical dependencies in the data and thereby allows performing correct statistical tests.

4 In this notation e is the idiosyncratic error and u is the unit-specific error (or random effect).

5 Obviously, the data structure of a classical panel is identical to the data structure produced by factorial surveys if each respondent has to rate multiple vignettes: Multiple observations of the dependent and independent variables from each respondent.

However, in addition to the issue of adequate statistical procedures, the SUMREG model has another advantage. It allows using the estimates of the random effects (u_i) for substantive interpretations. In the next section I will first lay out how to estimate the SUMREG model using generalized multilevel structural equation modeling (SEM). Then I will briefly discuss statistical tests, which are relatively straight forward, once the model has been presented, and finally I will introduce the idea of substantive interpretations of the random effects; this is, to conceptualize the unit-specific error components as latent variables.

Estimating the SUMREG Model Using Multilevel Structural Equation Modeling

Figure 1 presents the path diagram of a SUMREG model which includes j vignette dimensions as explanatory variables (X) and i ratings per vignette as dependent variables (Y). Each vignette dimension has a path to each of the dependent variables. The model furthermore includes random effects (u) for each dependent variable. These random effects (REs) are estimated at the level of the respondents. In other words, the path diagram shows a multilevel SEM in which the vignette dimensions and ratings are located at the first level (i.e. the vignette-level) and the REs are located at the second level (i.e. the respondent-level). The data structure for this model is in long format, i.e. the multiple vignettes asked per respondent each occupy a separate row, as it is typically done for standard multilevel modeling. What makes this model a SUR model are the correlations between the errors. More precisely, all idiosyncratic errors e are allowed to correlate with each other and all unit-specific errors u are allowed to correlate with each other.⁶

The introduction of respondent-level characteristics into that model can be done via the within-and-between formulation of multilevel models. Such a model is presented in Figure 2. It assumes that l respondent-level characteristics (Z) explain the between-unit variance in the dependent variables. As this variance is captured in the unit-specific REs, the respondent-level variables impact directly on these.⁷ This makes the formerly exogenous REs u endogenous variables, which are in Figure 2 indicated as η_i . The unexplained variance then is captured in the error term of these endogenous variables (u). In contrast to the variables measured at the vignette-level, the respondent-level variables may not all affect all dependent vari-

6 Given i dependent variables (i.e. ratings per vignette), the system includes $i*(i-1)/2$ covariances between the unit-specific error terms u as well as between the idiosyncratic error terms e .

7 There is also a different but equivalent formulation of that model, in which the respondent-level characteristics impact the dependent variables directly, i.e. a single-level formulation.

ables. As discussed above, the SUR approach allows that a subset of the explanatory variables affect only part of the dependent variables. In that case, the model would no longer be equivalent to a multivariate multilevel model.

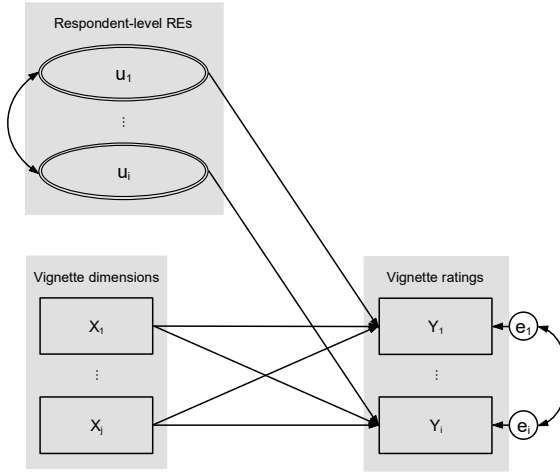


Figure 1 SUMREG Model with Vignette Dimensions

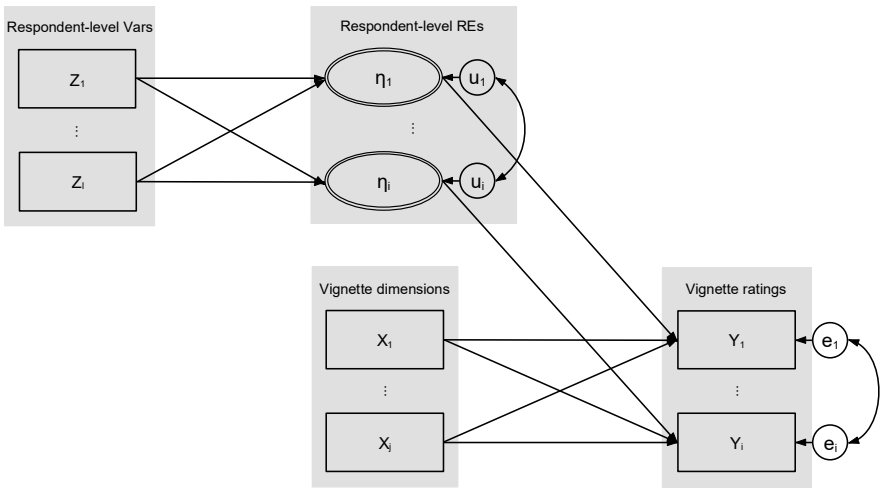


Figure 2 SUMREG Model with Vignette Dimensions and Respondent-level FEs

Comparing Parameters Within and Across Ratings

Factorial surveys with multiple ratings per vignette offer a variety of potential hypotheses tests. For example, we can ask whether a particular vignette characteristic has the same effects across ratings. We can also ask whether a set of vignette characteristics affects one dependent variable but not another, or whether the effect of a vignette characteristic on one rating is larger than on another rating, and so forth. Such hypotheses cannot be tested if the multiple ratings are modelled separately.

In general, there are two distinct ways of testing such hypotheses: We can either use a Wald-Test of linear hypothesis or we can compare a restricted and an unrestricted model using Likelihood-Ratio-Tests. Ultimately, these two tests are asymptotically equivalent (Engle, 1984) but, depending on the specific hypothesis to be tested, one or the other may present itself as more obvious. For example, if we want to test whether all explanatory variables have the same effect on each of the dependent variable, it seems more obvious to estimate an unrestricted model and compare it to a model with the appropriate restrictions via Likelihood-Ratio-Tests. If, on the other hand, we are interested in comparing two specific parameters, or testing one parameter against zero, the Wald-Test seems more appropriate.⁸

Conceptualizing the Random Effects as Latent Factors

From the viewpoint of standard multilevel modeling, the random effects u are merely error terms that capture the unexplained variance between the second-level clusters, i.e. respondents. In the language of panel data analysis, they would be described as unobserved heterogeneity (see Andress, Golsch, & Schmidt, 2013: 96f., for a discussion of the equivalence of multilevel and random effects panel data models). However, such random effects can be understood as latent variables (Skrondal & Rabe-Hesketh, 2004), which is the reason why SEM can be used to estimate multilevel models.

What exactly do these latent variables capture? As they are measured at the respondent-level, they reflect differences *between* respondents, independent of their reaction to specific vignettes. In other words, the REs reflect the tendency of respondents to select a specific response category independent of the varying stimuli. What stories can such REs tell?

⁸ Note that Likelihood-Ratio-Tests require re-estimating the model with the appropriate restrictions while Wald-Tests do not. Given the complexity of generalized multilevel SEMs, this process can take quite some time. If time is a scarce resource for you, you may prefer using Wald-Tests.

First, and foremost, when we compare the variance *between* units with the variance *within* units in an empty model, i.e. calculate the intra-class-correlation (ICC) coefficient, we can judge how much the respondents react to the experimental stimuli. If, for example, 90% of the total variance is between respondents, we could conclude that the vignette characteristics are generally not very effective. Such an analysis of the error variance can of course be done with simple multilevel models as well. However, the SUMREG model allows comparing the ICCs across the different ratings and thereby allows making statements about these differences. For example, it might be that one dependent variable reacts stronger to vignette characteristics than another.

Second, and this is specific to the SUMREG model, we can analyze the relationships between the REs of each dependent variable (rating). Such an analysis of the latent factor structure is directly included in the model, i.e. the variance-covariance matrix of the REs. The model gives us a clue as to how much the general tendencies of respondents to rate all vignettes similarly, are related across the different ratings. For example, we might see that some ratings are quite strongly related while others seem to be separate issues (compare Example 1 in Section 6.1).

If we allow ourselves to adapt more of the typical thinking of structural equation modelers, we see that we can even assume that two or more ratings are actually expressions of the same underlying latent variable. Thus, we could test a model in which all ratings are understood as being indicators of the same underlying issue, i.e. in which there is only a single RE instead of one per dependent variable.

Such a model is shown in Figure 3 and can be compared to a model with a separate RE for each dependent variable via Likelihood-Ratio-Tests (LR-Tests).

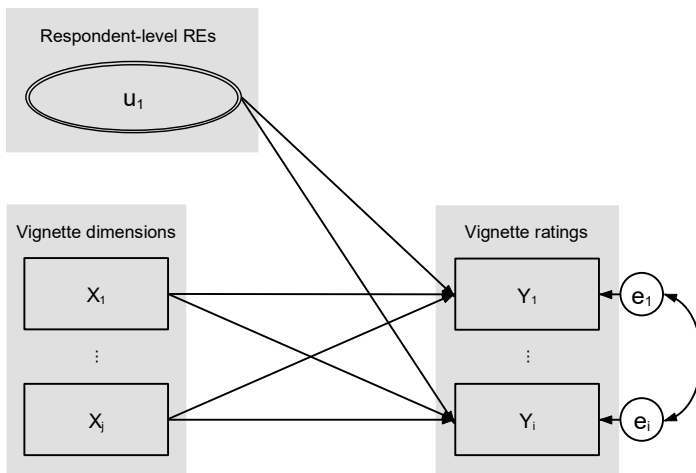


Figure 3 SUMREG Model with single RE for all Ratings

Depending on the number of ratings per vignette there are a variety of possible model specifications. If the factorial survey design consists of two ratings per vignette, there are just two possible models: A separate RE for each rating or one RE for both. If, however, the design includes more than two ratings per vignette, one might assume that some ratings share an underlying latent variable while others are separate issues, i.e. have their own REs. Similarly, we can obviously compare a model which assumes completely unrelated REs against a model which allows correlations between them. This would be an empirical test of the hypothesis that the issues are completely unrelated to each other.

Another nice feature of the SUMREG model is that we can predict the REs and analyze their joint distribution in detail. Such an analysis can be interesting in its own right but might make particular sense if the multiple ratings per vignette constitute something like a Guttman scale. Using predicted values of the REs in that case allows checking the consistency in response behavior. An example of such an analysis is shown below (Example 1.1).

Finally, a short note on the implied measurement model of the single REs is necessary. The SUMREG model in its unconstrained form, as in Figures 1 and 2, provides a RE for each rating. This effect is identified because each respondent rates multiple vignettes. As explained above, from the viewpoint of SEMs these REs can be understood as measurement models of latent factors. Then, of course, the question arises how the measurement coefficients (factor loadings) of that model look like. These parameters are not explicitly part of the model. As indicated above, the data for this kind of model is organized in long format with respondents each occupying as many rows in the data set as they have rated vignettes. The “measurement coefficients” of the REs therefore is the implicit coefficient of the respondent-level error term u_i , which is 1. Thus, each vignette is given the same weight in the “construction” of the respondent-level latent factors. This assumption might appear problematic but actually it is well justified. For each vignette respondents answer the same questions. Thus, the *wording* of a specific rating item is actually the same for each of its measurements. What varies between the measurements are just the descriptions on the vignettes.

Examples

All following analyses are performed using stata’s `gsem` command for generalized multilevel SEM (version 14.2). The data sets and do-files are provided in the online appendix of this paper (see footnote 1). I use two examples to demonstrate how the aforementioned modeling strategies and tests can be applied to real data. One example data set is from a factorial survey conducted in Germany in April 2015 (Czymara & Schmidt-Catran, 2016) and the other one from a factorial survey con-

Table 1 Description of Example Data Sets

	Czymara and Schmidt-Catran	Diehl et al.
<i>Sample</i>		
Full Sample (Respondent N)	1,283	1,432
Used Sample (Respondent N)	77	284
Fielded Vignettes per Respondent	14	4
Answered Vignettes per Respondent	14	3.98
Valid Vignette Ratings	1,078	1,131
<i>Vignette Characteristics</i>		
Vignette Dimensions	6	6
Total Vignette Levels	15	19
Vignette Universe	192	567
Ratings per Vignette	3	2
Points of Rating Scales	7	7

ducted in Switzerland between March and May 2014 (Diehl et al., 2018).⁹ In both cases I analyze a random sub-sample of the complete data, each of which includes about 1,000 unique vignette ratings.¹⁰

Both surveys deal with the impact of cultural and economic threats on the acceptance of immigrants. While the data by Czymara and Schmidt-Catran (2016) is based on a D-efficient design, in which all respondents received the same set of 14 vignettes (Dülmer, 2007: 385ff.), the data from Diehl et al. (2018) is based on a D-efficient sampling design, in which each respondent received a different subset of 4 vignettes (Dülmer, 2007: 384ff.). What both surveys have in common is that they generated data in which multiple vignettes are nested within respondents and respondents provided multiple ratings per vignette. Czymara and Schmidt-Catran (2016) use three ratings per vignette, while Diehl et al. (2018) use two ratings per vignette. Table 1 provides some information about both studies and the samples used for the following examples.

9 I like to give special thanks to Claudia Diehl, Katrin Auspurg and Thomas Hinz for providing their data.

10 I do so for two reasons: First, estimating these models is quite time consuming. By reducing the number of observations, I reduce the time needed for estimating and/or replicating my results. Second, I did not want to provide the full data from other authors.

This paper is certainly not the place for an extensive theoretical discussion but I will briefly summarize the central idea behind the two surveys: On the one hand, negative attitudes towards immigrants are assumed to be determined by natives' fear of the economic consequences of immigration (Facchini, Mayda, & Puglisi, 2013). On the other hand, scholars argue that natives fear the loss of their culture and therefore turn against immigrants (Hopkins, 2015). Factorial surveys are particularly well suited for this research area because they allow the simultaneous analysis of several determinants (i.e. cultural and economic threats) and minimize the risk of socially desirable answers (Wallander, 2009).

For the purpose of this paper it is sufficient to keep in mind that the vignettes cover economic and cultural characteristics of immigrants and that attitudes are expected to be particularly negative towards culturally more distinct immigrants. With regard to economic threats, the literature assumes that immigrants with higher skill levels are generally preferred because they should contribute more to the economic system in general (Hainmueller & Hiscox, 2007) but it has also been stated that natives fear competition on the job market (Facchini & Mayda, 2012).

Example 1: Czymara and Schmidt-Catran Data

In this factorial survey respondents were asked to rate vignettes with regard to three issues: Should the immigrant described on the vignette have the right to (1) live in Germany, (2) work in Germany, and (3) receive social benefits in Germany? Answers were measured on a 7-point scale, where higher values indicate willingness to grant the related right. The first step of the empirical analysis regards the factor structure of these three items.

Example 1.1: Analysis of Latent Factor Structure

Table 2 shows three empty models with a varying number of REs. M1 includes a separate RE for each rating (U1, U2, U3). All models allow correlations between the REs and also between the idiosyncratic errors. In Model M1 all of these correlations are highly significant, indicating that the SUMREG model is indeed justified. All models include an intercept for each rating, showing that acceptance of immigrants working and living in Germany is much higher (5.17 and 5.47 respectively) than acceptance of immigrants taking social benefits (3.88).¹¹ The variances of the REs reveal that between respondents, ratings vary much more with regard to the issue of social benefits than with regard to the other two issues. Thus, natives seem to have a stronger consensus over the issues of living and working in Germany

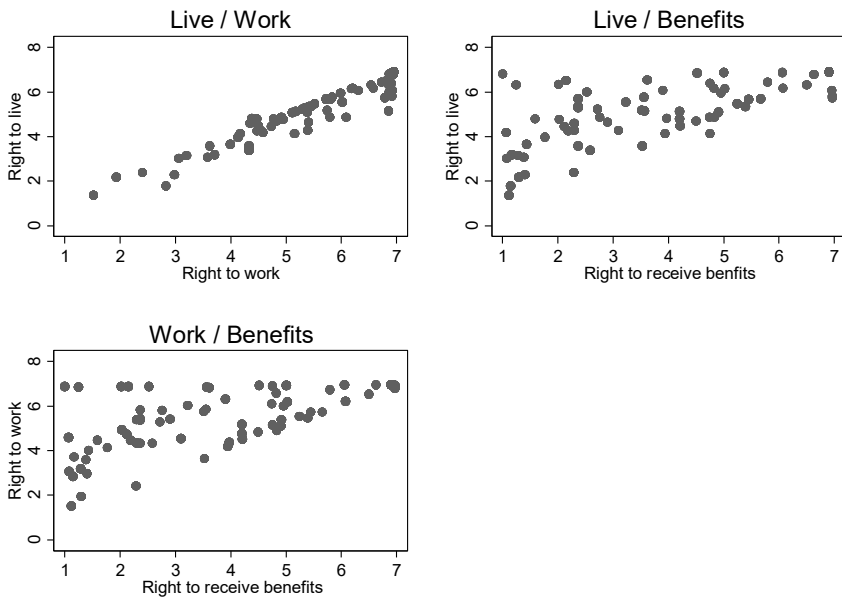
11 The „factor loadings“ of the REs (U1, U2, U3) are all 1 because each dependent variable has its own RE, which then by definition has to be 1 as it provides the anchor to scale the latent variable.

than over the issue of social benefits for immigrants. All of this, however, could also be seen from separate multilevel regressions. The covariances between the REs, in contrast, are unique to the SUMREG model. Note that Table 2 expresses covariances between the REs as correlations to allow for ease of comparison across pairs. While the issues of living and working in Germany are very closely related ($\text{Corr}(U1,U2)=.95$), the issue of social benefits seem to be less strongly associated with the other two ($\text{Corr}(U1,U3)=.68$, $\text{Corr}(U2,U3)=.66$).

Table 2 Empty SUMREG models - Example 1.1

	M1	M2	M3
<i>Live</i>			
U1	1.000 constr.	1.000 constr.	1.000 constr.
Intercept	5.171 ***	5.171 ***	5.171 ***
<i>Work</i>			
U2	1.000 constr.		
U1		0.958 ***	1.014 ***
Intercept	5.469 ***	5.469 ***	5.469 ***
<i>Benefits</i>			
U3	1.000 constr.		1.000 constr.
U1		2.891 ***	
Intercept	3.878 ***	3.878 ***	3.878 ***
<i>Variances and Covariances</i>			
Var(U1)	2.043 ***	0.425 ***	1.938 ***
Var(U2)	2.006 ***		
Var(U3)	3.837 ***		3.785 ***
Corr(U2,U1)	0.953 ***		
Corr(U3,U1)	0.682 ***		0.674 ***
Corr(U3,U2)	0.657 ***		
Var(e.live)	1.809 ***	3.427 ***	1.914 ***
Var(e.work)	1.576 ***	3.191 ***	1.588 ***
Var(e.benefits)	1.489 ***	1.775 ***	1.509 ***
Cov(e.work,e.live)	1.415 ***	2.938 ***	1.379 ***
Cov(e.benefits,e.live)	1.107 ***	1.787 ***	1.153 ***
Cov(e.benefits,e.work)	0.943 ***	1.588 ***	0.927 ***
<i>Statistics</i>			
Log-Likelihood	-4,699.36	-5,088.74	-4,790.21
LR-Tests		M2 vs. M1	M3 vs. M1
LR chi ²		778.75	181.68
Prob > chi ²		0.000	0.000

Notes: * p<.05, ** p<.01; *** p<.001 (two-sided tests).



Notes: Values are predicted as the sum of the intercept and the individual REs [BLUPs].

Figure 4 Predicted values from Model M1 – Example 1.1

Figure 4 presents the association of the three REs in more detail by means of scatter plots, using predicted values from Model M1 (Intercept + REs [BLUPs]). The first panel in Figure 4 shows that for almost all respondents the right to live and the right to work in Germany go together, i.e. they are on the diagonal of the plot. There is also a cluster of respondents which fully grant the right to work in Germany (7 on the x-axis) but do not grant the right to live in Germany to the same extent, i.e. they are below the diagonal. Following these insights one could categorize such response behavior as inconsistent and decide how to treat these cases.¹²

The second and third panels in Figure 4 look quite similar, with all respondents being on or above the diagonal, indicating that a large share of respondents tends to grant the right to live (panel 2) or work (panel 3) in Germany to a larger extent than the right to receive social benefits. Panel 2 again reveals two respondents that provide inconsistent answers, granting the right to receive benefits but to a lesser degree the right to live in Germany.

12 This paper is not the place for a detailed discussion of such issues but there are a number of alternatives: One could simply recognize that some respondents give inconsistent answers and go on with the analysis or one could exclude these respondents from the analysis. One could also think of using the SUMREG model during the pre-test phase of a factorial survey and take such a result as an indicator that the vignettes and the instructions may need to be redesigned.

Models M2 and M3 in Table 2 test whether the three separate REs from model M1 can be replaced by shared factors. Model M2 includes one RE for all three dependent variables (U1). The model does fit the data significantly worse than model M1 (LR-Test: $p < .0001$) and therefore we can conclude that the three ratings are not expressions of the same underlying latent factor. This result is not surprising given the graphical evidence from Figure 4. The variance and covariance parameters in model M2 are not of great interest but readers should note that the factor loadings, which have all been 1 in model M1 are now allowed to vary across ratings, making them true factor loadings in this model. The first factor loading (live) is still 1 as it provides the anchor to scale the latent variable.

Model M3 assumes that the issues of living and working in Germany share one underlying latent factor (U1) while the issue of receiving benefits has its own RE (U3). Given the evidence from Figure 4 this seems like a reasonable assumption, but the model does not hold against model M1 (LR-Test: $p < .0001$). Thus, we can conclude that this data is best modeled with a separate RE for each rating: Living, working and receiving benefits in a host society seem to be separate issues, where respondents can show various combinations of positive and negative attitudes. Such a conclusion could not be tested without the SUMREG model.

Example 1.2: Analyzing Fixed Effects

Table 3 presents two models in which the vignette-level effects have been added to the fixed part of the equations. Both models include a separate RE for each of the ratings, following the evidence from Models M1, M2 and M3. Model M4 estimates separate fixed effects for each of the three dependent variables, while Model M5 constrains them to be equal across all three ratings. The LR-Test comparing both models indicated that Model M5 does fit the data significantly worse than Model M4. We can therefore conclude that the vignette characteristics' effects are not identical across the three dependent variables, at least if one tests all of them simultaneously. Again, such a conclusion requires the SUMREG model for a correct statistical test.

Substantially, the results show that there is no effect of an immigrants' gender or country of origin on her or his acceptance. Immigrants with higher education and good language skills are preferred over those with lower education and bad language skills. Muslim immigrants are less accepted than immigrants who are Christians or do not have a religious denomination, but this is only significant for the right to live in Germany not for the other two issues. The strongest effect, however, is a person's motivational reason for immigration. Immigrants that have a job in prospect are much more welcome than those who come for economic reasons but without any economic prospects. Immigrants who flee from political persecution are by far the most accepted group.

Table 3 Adding vignette-level covariates – Example 1.2

	M4			M5
	Live	Work	Benefits	Live/Work/ Benefits
Gender (Ref. = Female)				
Male	0.029	0.078	-0.011	0.027
Country of Origin (Ref. = Lebanon)				
France	0.135	0.144	0.152	0.147
Kenya	-0.074	-0.086	-0.043	-0.063
Reason for Migr. (Ref. =better live)				
Political Persecution	1.420 ***	1.045***	1.333 ***	1.239 ***
Job	0.939 ***	0.813***	0.632 ***	0.736 ***
Education (Ref. = low education)				
University	0.337 ***	0.307***	0.193 **	0.253 ***
Language skills (Ref. = none)				
Good	0.471 ***	0.420***	0.271 ***	0.350 ***
Religion (Ref. = no Religion)				
Christ	0.055	0.042	0.055	0.050
Muslim	-0.230 **	-0.138	-0.060	-0.110
U	1.000 constr.	1.000 constr.	1.000 constr.	1.000
Intercept Live	4.020 ***			4.204 ***
Intercept Work		4.463 ***		4.503 ***
Intercept Benefits			2.985 ***	2.911 ***
<i>Variances and Covariances</i>				
Var(U1)		2.073 ***		2.071 ***
Var(U2)		2.025 ***		2.023 ***
Var(U3)		3.857 ***		3.856 ***
Corr(U2,U1)		0.953 ***		0.954 ***
Corr(U3,U1)		0.683 ***		0.684 ***
Corr(U3,U2)		0.658 ***		0.659 ***
Var(e.live)		1.390 ***		1.410 ***
Var(e.work)		1.308 ***		1.326 ***
Var(e.benefits)		1.209 ***		1.221 ***
Cov(e.work,e.live)		1.085 ***		1.091 ***
Cov(e.benefits,e.live)		0.782 ***		0.774 ***
Cov(e.benefits,e.work)		0.698 ***		0.684***

	M4			M5
	Live	Work	Benefits	Live/Work/ Benefits
<i>Statistics</i>				
Log-Likelihood		-4528.1653		-4577.7834
LR-Tests				M4 vs. M3
LR chi2				99.24
Prob > chi2				0.000

Notes: * $p < .05$, ** $p < .01$, *** $p < .001$ (two-sided tests).

As discussed above, the SUMREG model allows performing statistically correct tests across the multiple ratings. One such test is the LR-Test comparing the two models presented in Table 3. In order to compare single coefficients or test a few selected parameters, the Wald-Test seems to present itself since it does not require re-estimating the model. For example, we could hypothesize that an immigrant's education is more important for work-related issues than for the general right to live in Germany. Or, vice versa, we could hypothesize that an immigrant's qualification is equally important for all three issues. The corresponding test on the coefficients estimated in Model 4 indicates that the effect is indeed independent of the specific issue ($\text{Chi}^2=5.20$, $p=.074$).

We might wonder whether being a Muslim matters more for an immigrant's general acceptance (right to live, coef. = -0.230) than for granting her or him the right to receive social benefits (coef. = $-.060$). Using a Wald-Test we can check whether the effect of being a Muslim on the right to live is significantly stronger than the effect on the right to receive social benefits. The test reveals that it actually is: $\text{Chi}^2=4.96$, $p=.026$.¹³ As these examples show, hypotheses about differences and commonalities across the ratings can be theoretically meaningful. In order to test such hypotheses, the SUMREG model is required, as a separate modeling of the ratings does not allow to perform such tests.

Another interesting perspective opened by the SUMREG model is related to the covariation of the idiosyncratic error terms. In an empty model (compare Model M1 in Table 2), the covariance between these error terms reflect not only correla-

13 This is an example where a naive statistical test based on separate multilevel regression models would give a different result: When testing the effect of Muslim on the right to work against the numerical value of the coefficient in the model of social benefits (-0.60), the test indicates a non-significant difference ($\text{Chi}^2=3.69.87$, $p=.055$). The univariate multilevel models used for this naive and incorrect (!) test can be found in the online appendix (see footnote 1): Univariate M4, Table OA2.

tions between the idiosyncrasies of the ratings but also their joint variation due to the treatments on the vignettes, i.e. the vignette-level effects. Once the vignette characteristics are controlled, this “explained” part of the covariance is removed from the random part of the model and the remaining covariances of the residuals indicate “unexplained” covariance between the idiosyncratic error terms. If this unexplained covariance remains substantial, we might take this as an indicator of problematic response behavior. For example, respondents may have thought only about the first rating and then simply selected the same scale points for the remaining ratings. In the example above, the covariation of the three idiosyncratic error terms (live, work, social benefits) has been reduced by 23%, 29% and 26%, respectively, when comparing the empty Model M1 and Model M4. Thus, a substantial amount of covariance is left after accounting for the vignette-level effects.

Example 1.3: A True Seemingly Unrelated Regression Model

While the SUMREG models include some parameters that are obviously missing in a univariate approach (i.e. the covariances of REs and idiosyncratic errors across ratings), the models presented so far provide estimates that are identical to those from simple univariate multilevel models (compare Table OA2 in the online appendix, see footnote 13). In this sense, the multivariate approach of the SUMREG model simply adds the potential to statistically compare coefficients across equations. However, as indicated above, the estimates from seemingly unrelated regressions differ from univariate estimates if the set of predictors varies between equations. In that case the SUR approach is no longer equivalent to a multivariate regression model. Such a *true* seemingly unrelated regression model benefits from a gain in efficiency resulting from the “zero restrictions” implied by the model specification (compare Zellner 1962: 353 f.).

Table 4 presents two SUMREG models which include, in addition to the vignette-level effects, the respondent-level variable *education*. In Model M6 education is included in each of the three equations while in Model M7 it influences only the right to work. The decision to assume an effect of education only on the work-related rating, as in Model M7, may be theoretically motivated; reflecting the idea that economic characteristics should matter most for employment-related issues, where competition on the labor market could be important, and less for the general acceptance or immigrants’ deservingness of social assistance. A comparison of Models M6 and M7 illustrates the gain in efficiency due to the seemingly unrelated regression approach: While the effect of education is not significant on any of the three dependent variables in Model M6, it is significant in Model M7. The standard errors of the education effect are more than three times smaller in the latter model. Of course, Model M7 should be tested against M6 before one selects it as the better model. An LR-Test comparing the two models indicates that the model fit of them is not significantly different ($p=.94$). Thus, from a model-fit-perspective one could

Table 4 Adding respondent-level covariates - Example 1.3

	M6			M7		
	Live	Work	Benefits	Live	Work	Benefits
<i>Vignette-level Effects</i>						
Gender (Ref. = Female)						
Male	0.046 (0.075)	0.081 (0.074)	-0.015 (0.070)	0.046 (0.075)	0.081 (0.074)	-0.015 (0.070)
Country of origin (Ref. = Lebanon)						
France	0.144 (0.108)	0.149 (0.106)	0.172 (0.100)	0.144 (0.108)	0.149 (0.106)	0.172 (0.100)
Kenya	-0.065 (0.092)	-0.090 (0.090)	-0.046 (0.085)	-0.065 (0.092)	-0.090 (0.090)	-0.046 (0.085)
Reason for coming (Ref. = better live)						
Political Persecution	1.466 (0.108) ***	1.087 (0.106) ***	1.354 (0.100) ***	1.466 (0.108) ***	1.087 (0.106) ***	1.354 (0.100) ***
Job	0.979 (0.092) ***	0.846 (0.090) ***	0.670 (0.085) ***	0.979 (0.092) ***	0.846 (0.090) ***	0.670 (0.085) ***
Education (Ref. = low education)						
University	0.338 (0.075) ***	0.320 (0.074) ***	0.206 (0.070) **	0.338 (0.075) ***	0.320 (0.074) ***	0.206 (0.070) **
Language skills (Ref. = none)						
Good	0.478 (0.075) ***	0.437 (0.074) ***	0.278 (0.070) ***	0.478 (0.075) ***	0.437 (0.074) ***	0.278 (0.070) ***
Religion (Ref. = no Religion)						
Christ	0.068 (0.098)	0.044 (0.095)	0.051 (0.091)	0.068 (0.098)	0.044 (0.095)	0.051 (0.091)
Muslim	-0.202 (0.091) *	-0.144 (0.089)	-0.057 (0.084)	-0.202 (0.091) *	-0.144 (0.089)	-0.057 (0.084)
U	1.000 constr.	1.000 constr.	1.000 constr.	1.000 constr.	1.000 constr.	1.000 constr.
Intercept	3.878 (0.226) ***	4.273 (0.222) ***	2.816 (0.277) ***	3.898 (0.211)	4.291 (0.208)	2.856 (0.255)

	M6			M7		
	Live	Work	Benefits	Live	Work	Benefits
<i>Respondent-level Effects</i>						
Education (Ref. = low education)						
University	0.105 (0.436)	0.460 (0.427)	0.210 (0.580)	0.000 constr.	0.363 (0.131) **	0.000 constr.
<i>Variances and Covariances</i>						
Var(U1)		2.060 (0.355) ***			2.062 (0.356) ***	
Var(U2)		1.976 (0.341) ***			1.978 (0.341) ***	
Var(U3)		3.731 (0.628) ***			3.738 (0.629) ***	
Corr(U2,U1)		0.959 (0.168) ***			0.959 (0.168) ***	
Corr(U3,U1)		0.680 (0.145) ***			0.680 (0.145) ***	
Corr(U3,U2)		0.645 (0.143) ***			0.645 (0.143) ***	
Var(e.live)		1.409 (0.064) ***			1.409 (0.064) ***	
Var(e.work)		1.349 (0.062) ***			1.349 (0.062) ***	
Var(e.benefits)		1.219 (0.056) ***			1.219 (0.056) ***	
Cov(e.work,e.live)		1.120 (0.057) ***			1.120 (0.057) ***	
Cov(e.benefits,e.live)		0.811 (0.050) ***			0.811 (0.050) ***	
Cov(e.benefits,e.work)		0.720 (0.047) ***			0.720 (0.047) ***	

Notes: * p<.05, ** p<.01, *** p<.001 (two-sided tests). The models presented in this table differ in the number of observations (n=1,036) from the previous models (n=1,078) because some respondents (N=3) have missing values on education. This explains the difference in vignette-level effects.

select the more parsimonious model (M7) and thereby harvest the efficiency gain from the SUMREG model.

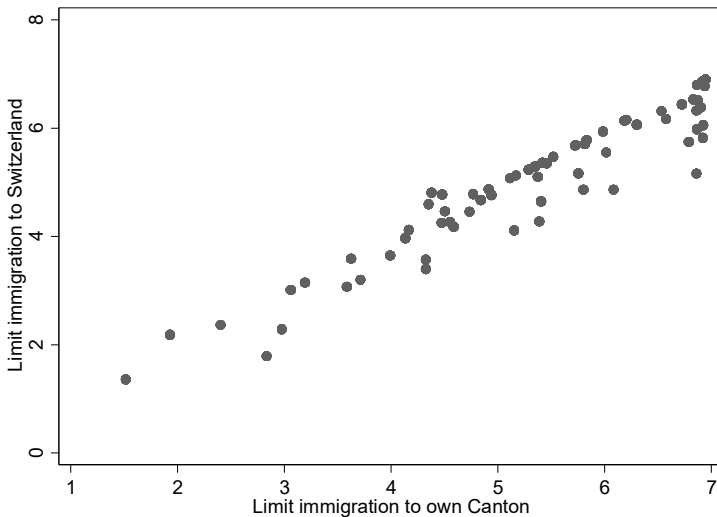
Example 2: Diehl et al. Data

In the study of Diehl et al. (2018) respondents were asked to provide two ratings per vignette: Should immigration from the described group be limited (1) to Switzerland in general and (2) to the respondent's own canton? Each rating was done on a 7-point scale where higher values indicate a desire to limit migration.

Example 2.1: Analysis of Latent Factor Structure

Intuitively these two ratings appear to have more in common than the three issues addressed in the former example, so we may expect to find them to be expressions of one latent factor, and this is exactly what an analysis of the underlying factor structure reveals, thereby providing a good counter example to the analysis above.

Table 5 presents two empty models. Model M1 includes a separate RE for each rating and Model M2 assumes that both ratings share one underlying latent factor. The correlation between the two REs in Model M1 is almost perfect (.99) and the LR-Test comparing the two models indicates that one can indeed model the two ratings as being expressions of the same underlying latent factor. Figure 5 presents the relationship between the two random components of Model M1 graphically.



Notes: Values are predicted as the sum of the intercept and the individual REs [BLUPs].

Figure 5 Predicted values from Model M1 – Example 2.1

Table 5 Empty SUMREG models - Example 2.1

	M1	M2
Switzerland		
U1	1.000 constr.	1.000 constr.
Intercept	3.671 ***	3.671 ***
Own Canton		
U2	1.000 constr.	
U1		1.013 ***
Intercept	3.751 ***	3.751 ***
<i>Variances and Covariances</i>		
Var(U1)	1.790 ***	1.788 ***
Var(U2)	1.841 ***	
Corr(U2,U1)	0.995 ***	
Var(e.switzerland)	2.194 ***	2.197 ***
Var(e.owncanton)	2.344 ***	2.351 ***
Cov(e.switzerland,e.owncanton)	2.075 ***	2.071 ***
<i>Statistics</i>		
Log-Likelihood	-3,337.73	-3,339.21
LR-Tests		M2 vs. M1
LR chi ²		2.95
Prob > chi ²		0.086

Notes: * p<.05, ** p<.01, *** p<.001 (two-sided tests).

Example 2.2: Analyzing Fixed Effects

Table 6 presents the results of two SUMREG models that include the vignette-level effects. According to the results from above (Example 2.1), both models assume that the two ratings are expressions of the same underlying RE. What differs between the two models is that Model M3 estimates separate fixed effects on the two ratings, while Model M4 constraints the effects to be equal. The LR-Test comparing the two models reveals that Model M3 does not have a significantly better fit and we can therefore conclude that the vignette dimensions affect both ratings equally.

Substantively the results show that immigrants from countries that are culturally more distant from Switzerland (Romania and Croatia) are less accepted. Immigrants with higher education are preferred over immigrants with basic education. Intended duration of stay does not have a significant effect. Immigrants that intend to find jobs for which no Swiss people are available are more accepted than immigrants who look for jobs that also Swiss people are looking for. Respondents have

Table 6 Adding vignette-level covariates to the model – Example 2.2

	M3		M4
	Switzerland	Own Canton	Switzerl./ Own Cant.
Country of origin (Ref. = Germany)			
France	0.305 *	0.376 *	0.323 *
Italy	-0.074	-0.069	-0.072
Norway	0.111	0.106	0.110
Romania	1.053 ***	0.954 ***	1.027 ***
Croatia	0.726 ***	0.716 ***	0.723 ***
Education (Ref. = University)			
Basic Education	0.322 ***	0.312 ***	0.319 ***
Intended duration of stay (Ref. = for ever)			
Several Years	0.022	-0.070	-0.002
One Year	-0.004	-0.047	-0.015
Swiss people available for job (Ref. = no)			
Yes	0.735 ***	0.668 ***	0.718 ***
Language skills (Ref. = German and French)			
No German and no French	1.021 ***	1.088 ***	1.039 ***
French but no German	0.514 ***	0.559 ***	0.526 ***
German but no French	0.249 *	0.310 *	0.266 *
Culture (Ref. = willing to adapt)			
Not willing to adapt	0.751 ***	0.720 ***	0.742 ***
No information	0.418 *	0.426 *	0.415 *
U1	1.000 constr.	1.015 ***	1.000 constr.
Intercept	1.942 ***	2.080 ***	1.960 ***
<i>Variances and Covariances</i>			
Var(U1)	1.819 ***		1.818 ***
Var(e.switzerland)	1.653 ***		1.654 ***
Var(e.owncanton)	1.831 ***		1.833 ***
Cov(e.switzerland,e.owncanton)	1.542 ***		1.541 ***
<i>Statistics</i>			
Log-Likelihood	-3,203.86		-3,211.75
LR-Tests			M4 vs. M3
LR chi ²			15.78
Prob > chi ²			0.327

Notes: * p<.05, ** p<.01, *** p<.001 (two-sided tests).

a preference for immigrants that speak at least one of the official languages or even better speak German and French. Immigrants that are willing to adapt to the Swiss culture are most accepted, while those who do not want to adopt are least accepted.

In sum, the survey of Diehl et al. (2018) provides an example of a vignette study in which (1.) the multiple ratings can be understood as expressions of the same underlying latent concept and (2.) the effects of the vignette characteristics are the same across the multiple ratings. The SUMREG models therefore allows for a very parsimonious parameterization of the model explaining the data (Model M4).

Summary and Discussion

In this paper I proposed a modeling approach for factorial surveys with multiple ratings per vignette. As shown in a literature review, factorial surveys with multiple ratings are not uncommon. The SUMREG model estimates the equations for each of these dependent variables simultaneously, while allowing the error terms of the equations to correlate with each other. This allows for a statistically correct comparison of coefficients across ratings via LR- or Wald-Tests. If expected differences of coefficients can be derived from theoretical considerations, the SUMREG model allows for a more encompassing test of these theories.

The model, furthermore, allows conceptualization of the REs as latent factors and analyses of the latent factor structure underlying the ratings. Due to the use of multilevel SEM the procedure allows estimation of models which assume that all (or a subset) of the ratings are expressions of the same underlying latent factors. In case that such a model fits the data, as in Example 2.1, SUMREG allows a more parsimonious model specification. Additionally, one can restrict the vignette-level effects to be equal across equations. If such a model holds, as in Example 2.2, the SUMREG model allows for a very parsimonious model specification.

The proposed model can be applied to any data from factorial surveys that (1.) include multiple ratings per vignette (at least 2) and (2.) multiple vignettes per respondent (at least 2). Without the latter, the REs are not identified. The model would then reduce to a simple SUR model, which still has the benefit of providing correct comparisons of coefficients across the ratings. The model could be extended by the inclusion of random slopes, which would allow for the estimation of respondent-specific vignette-level effects. This requires, however, a sufficiently large number of vignettes per respondent, but future work should consider such a model extension.

References

- Andress, H.-J., Golsch, K., & Schmidt, A. W. (2013). *Applied Panel Data Analysis for Economic and Social Surveys*. Berlin, Heidelberg: Springer.
- Atzmüller, C., & Steiner, P. M. (2010). Experimental vignette studies in survey research. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 6, 128-138.
- Auspurg, K., Hinz, T., Sauer, C., & Liebig, S. (2015). The factorial survey as a method for measuring sensitive issues. In U. Engel, B. Jann, P. Lynn, A. Scherpenzeel & P. Sturgis (Eds.), *Improving Survey Methods: Lessons From Recent Research* (pp. 137-149). New York: Routledge.
- Baltagi, B. H. (1980). On Seemingly Unrelated Regressions with Error-Components. *Econometrica*, 48(6), 1547-1551. doi: 10.2307/1912824
- Czymara, C. S., & Schmidt-Catran, A. W. (2016). Wer ist in Deutschland willkommen? *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 68(2), 193-227.
- Diehl, C., Auspurg, K., & Hinz, T. (2018). Do Perceptions of Economic and Cultural Threat Explain Preferences for Immigration Control in Switzerland? *Accepted for publication in: Swiss Journal of Sociology*, 44(1).
- Dülmer, H. (2007). Experimental Plans in Factorial Surveys. *Sociological Methods & Research*, 35, 382-409.
- Dülmer, H. (2016). The Factorial Survey Design Selection and its Impact on Reliability and Internal Validity. *Sociological Methods & Research*, 45(2), 304-347.
- Engle, R. F. (1984). Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. *Handbook of econometrics*, 2, 775-826.
- Facchini, G., & Mayda, A. M. (2012). Individual Attitudes Towards Skilled Migration: An Empirical Analysis Across Countries. *The World Economy*, 35(2), 183-196.
- Facchini, G., Mayda, A. M., & Puglisi, R. (2013). Individual Attitudes towards Immigration -- Economic vs. Non-economic Determinants. In G. P. Freeman (Ed.), *Immigration and Public Opinion in Liberal Democracies* (pp. 129-157). New York: Routledge.
- Hainmueller, J., & Hiscox, M. J. (2007). Educated preferences: Explaining attitudes toward immigration in Europe. *International Organization*, 61, 399-442.
- Hainmueller, J., Hopkins, D. J., & Yamamoto, T. (2014). Causal inference in conjoint analysis: understanding multidimensional choices via stated preference experiments. *Political Analysis*, 22, 1-30.
- Harell, A., Soroka, S., Iyengar, S., & Valentino, N. (2012). The Impact of Economic and Cultural Cues on Support for Immigration in Canada and the United States. *Canadian Journal of Political Science-Revue Canadienne De Science Politique*, 45(3), 499-530. doi: 10.1017/S0008423912000698
- Hopkins, D. J. (2015). The upside of accents: language, inter-group difference, and attitudes toward immigration. *British Journal of Political Science*, 45, 531-557.
- Hox, J. J., Kreft, I. G. G., & Hermkens, P. L. J. (1991). The Analysis of Factorial Surveys. *Sociological Methods & Research*, 19(4), 493-510. doi:10.1177/0049124191019004003
- Jasso, G. (2006). Factorial Survey Methods for Studying Beliefs and Judgments. *Sociological Methods & Research*, 34(3), 334-423. doi:10.1177/0049124105283121
- Jasso, G. (2019). Factorial Survey. In P. Atkinson, S. Delamont, A. Cernat, J.W. Sakshaug & R.A. Williams (Eds.), *SAGE Research Methods Foundations*. doi: 10.4135/9781526421036888176

- Rossi, P., & Nock, S. (Eds.). (1982). *Measuring Social Judgments: The Factorial Survey Approach*. Newbury: Sage Publications.
- Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: multilevel, longitudinal, and structural equation models*. Boca Raton: Chapman & Hall/CRC.
- Wallander, L. (2009). 25 years of factorial surveys in sociology: A review. *Social Science Research*, 38, 505-520.
- Weinberg, J. D., Freese, J., & McElhattan, D. (2014). Comparing Data Characteristics and Results of an Online Factorial Survey between a Population-Based and a Crowdsourced-Recruited Sample. *Sociological Science*, 1, 292-310.
- Zellner, A. (1962). An efficient method of estimating seemingly unrelated regression equations and tests for aggregation bias. *Journal of the American Statistical Association*, 57, 348-368.

Information for Authors

Methods, data, analyses (mda) publishes research on all questions important to quantitative methods, with a special emphasis on survey methodology. In spite of this focus we welcome contributions on other methodological aspects.

Manuscripts that have already been published elsewhere or are simultaneously submitted to other journals will not be considered. As a rule we do not restrict authors' rights. All rights remain with the author, and articles in mda are published under the CC-BY open-access license.

Mda aims for a quick peer-review process. All papers submitted to mda will first be screened by the editors for general suitability and then double-blindly reviewed by at least two reviewers. The decision on publication is made by the editors based on the reviews. The editorial team will contact the authors by email with the result at the latest eight weeks after submission; if the reviews have not been received by then, we provide a status update with a new target date.

When preparing a paper for submission, please consider the following guidelines:

- Please submit your manuscript via www.mda.gesis.org.
- The total length of the manuscript shall not exceed 10.000 words.
- Manuscripts should...
 - be written in English, using American English spelling. Please use correct grammar and punctuation. Non-native English speakers should consider a professional language editing prior to publication.
 - be typed in a 12 pt Roman font, double-spaced throughout.
 - be submitted as MS Word documents.
 - start with a cover page containing the title of the paper and contact details / affiliations of the authors, but be anonymized for review otherwise.
 - should be anonymized (“blinded”) for review.
- Please also send us an abstract of your paper (approx. 300 words), a front page with a brief biographical note (no longer than 250 words as supplementary file), and a list of 5-7 keywords for your paper.
- Acceptable formats for Graphics are
 - pdf
 - jpg (uncompressed, high quality)
- Please ensure a resolution of at least 300 dpi and take care to send high-quality graphics. Line art images should have a resolution of 500-1000 dpi. Please note that we cannot print color images.
- The type area of our journal is 11.5 cm (width) x 18.5 cm (height). Please consider this when producing tables or graphics.
- Footnotes should be used sparingly.
- Please number the headings of your article. Doing so will make the work of the layout editor easier.

- If your text includes formulas we would like to ask you to upload your text also as a PDF, additional to the Word document.
- By submitting a paper to mda the authors agree to make data and program routines available for purposes of replication.
- Response rates in research papers or research notes, where population surveys are analyzed, should be calculated according to AAPOR standard definitions.
- Please follow the APA guideline when structuring and formatting your work.

When preparing in-text references and the list of references please also follow the APA guidelines:

Entire Book:

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: John Wiley & Sons.

Journal Article (with DOI):

Klimoski, R., & Palmer, S. (1993). The ADA and the hiring process in organizations. *Consulting Psychology Journal: Practice and Research*, 45(2), 10-36. doi:10.1037/1061-4087.45.2.10

Journal Article (without DOI):

Abraham, K. G., Helms, S., & Presser, S. (2009). How social processes distort measurement: The impact of survey nonresponse on estimates of volunteer work in the United States. *American Journal of Sociology*, 114(4), 1129-1165.

Chapter in an Edited Book:

Dixon, J., & Tucker, C. (2010). Survey nonresponse. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research*. Second Edition (pp. 593-630). Bingley: Emerald.

Internet Source (without DOI):

Lewis, O., & Redish, L. (2011). *Native American tribes of Wisconsin*. Retrieved April 19, 2012, from the Native Languages of the Americas website: www.native-languages.org/wisconsin.htm

For more information, please consult the Publication Manual of the American Psychological Association (Sixth ed.).

gesis

Leibniz Institute for the Social Sciences

ISSN 1864-6956 (Print)

ISSN 2190-4936 (Online)

© of the compilation GESIS, Mannheim, July 2022