



Universiteit Antwerpen
**CSB | Centrum voor Sociaal
Beleid Herman Deleeck**

Veerle Van Loon and Koen Decanq

**Using a factorial survey to estimate the relative
importance of well-being dimensions according
to older people: insights from a repeated survey
experiment in Flanders**

WORKING PAPER

No. 21/09

September 2021



University of Antwerp
Herman Deleeck Centre for Social Policy
<https://www.uantwerpen.be/en/research-groups/csb/>



Using a factorial survey to estimate the relative importance of well-being dimensions according to older people: insights from a repeated survey experiment in Flanders

Veerle Van Loon¹ and Koen Decancq²

¹ Herman Deleeck Centre for Social Policy (University of Antwerp) veerle.vanloon@uantwerpen.be

² Herman Deleeck Centre for Social Policy (University of Antwerp); Department of Economics (KU Leuven); Centre for Philosophy of Natural and Social Science (London School of Economics) and CORE (UCLouvain) koen.decancq@uantwerpen.be

Working Paper No. 21/09

September 2021

Abstract

In this paper, we investigated the potential of a factorial survey to estimate the relative importance of the well-being dimensions of health, income, social relations, leisure and religion or spirituality, according to the views of older people. For this purpose, a factorial survey was implemented in a longitudinal online survey among 800 older adults in Flanders (the Flemish region of Belgium). The potential of the factorial survey was explored in two ways. First, we performed several within-sample test-retests to investigate the consistency of the estimated relative importance weights over time (i.e., temporal reliability). Second, we tested the feasibility of the factorial survey for the target population by studying two indicators of cognitive load: response time and response consistency. Overall, we found evidence that the factorial survey works well among a sample of older people. Health, income and social relations were the most important well-being dimensions – followed by leisure and engaging activities. Religion or spirituality appeared to be rather unimportant. The results proved to be consistent in most of the test-retest analyses. In addition, we found that respondents were able to cope well with the complexity of the factorial survey and produced a high level of response consistency within an acceptable amount of response time.

Keywords: older people, relative importance, well-being dimensions, factorial survey, feasibility, temporal reliability

INTRODUCTION

Over the next three decades, the number of older people is expected to double worldwide, reaching over 1.5 billion in 2050 (United Nations, 2020). This rapid aging and the associated pressure on public health care systems and funding has intensified international interest in the promotion of well-being at an older age (Pruchno, 2015). In this vein, the World Health Organization has recently declared a Decade of Healthy Aging (2020-2030), with its action plan aiming to improve the well-being of older people by putting their experiences and expertise at the center.

However, there is currently no firm consensus as to what well-being at an older age exactly entails (Fernández-Ballesteros, 2011). Indeed, in-depth interviews with older people revealed a multidimensional understanding of the concept and hence a divergence from dominant biomedical conceptualizations (e.g., the model of successful aging of Rowe and Kahn (1997)). While the majority of biomedical models include physiological factors such as disease, disabilities and physical functioning (Cosco et al., 2014; Depp & Jeste, 2006), only a few include dimensions beyond health that could be of value to older people, such as daily functioning, social engagement and personal resources (Cosco et al., 2013; Hung et al., 2010; Jopp et al., 2015; Phelan et al., 2004). This discrepancy was further evidenced by the fact that many older people considered themselves to be aging successfully without meeting the expert or researcher-defined criteria (Strawbridge et al., 2002).

Against this backdrop, several scholars have argued that there is a need to overcome what they consider is a paternalistic approach, in which experts or researchers define the meaning of well-being at an older age (Bowling & Dieppe, 2005; Decancq & Michiels, 2019; Phelan & Larson, 2002; Whitley et al., 2020). According to Bowling and Dieppe (2005: 1548-1550), there is “little point in developing policy goals if elderly people do not regard them as relevant” and any definition of successful aging therefore “needs to include elements that matter to elderly people.” There are several methods that include lay views in the definition of well-being (see, e.g., Cosco et al., 2013; Gabriel & Bowling, 2004; Jopp et al., 2015; Phelan et al., 2004; Pruchno et al., 2010). Existing studies use, for instance, open-ended questions such as: “How would you define successful aging?,” or ask older people to rate or rank life dimensions according to their importance (Bowling & Gabriel, 2004; Hackert et al., 2019; Hsieh, 2005; Jopp et al., 2015;

Phelan et al., 2004, Wilhelmson et al., 2005). Hsieh (2005), for example, found that health is the most important life dimension among older people in the US, followed by family life, religion, friendships, financial situation, spare time, neighborhood, and work.

A consequence of considering well-being at an older age as a multidimensional phenomenon that should reflect the view of older people is that comparisons of well-being become more complex. Indeed, how can we compare the well-being of an older person who is in better health with an older person who is in a better financial situation? To answer this and similar questions, it is necessary to assign relative importance weights to the different dimensions so that performance in the different well-being dimensions can be traded-off against each other.¹ Open-ended questions and direct rating or ranking of the dimensions are not very useful for this purpose as they provide little to no information about the intensity of the relative importance of the dimensions expressed and, hence, about the trade-offs between them.

In this study, we present the factorial survey as an alternative approach to deriving the views of older people on the relative importance of well-being dimensions. The factorial survey is an experimental method in which respondents are asked to rate several hypothetical descriptions of objects or situations (called vignettes) (Atzmüller & Steiner, 2010; Auspurg & Hinz, 2015). This approach provides information about the intensity of the trade-offs between dimensions and is, to the best of our knowledge, novel in the context of research on well-being at an older age. Although the factorial survey has been applied in a wide range of academic disciplines to address human judgments (for an overview, see Wallander, 2009), we found only one other study that used it for the similar purpose of well-being measurement, but among a different target group (namely the general population) (Whitley et al., 2020).

It remains an open question whether factorial surveys can be used in a reliable way when the target population is older people (Teti et al., 2016). Because of their multifactorial design and sometimes complex vignette descriptions, factorial surveys may put greater cognitive burden on respondents than standard survey questions. Moreover, it is well documented that cognitive capacity declines with age (Sauer et al., 2011). Although previous studies implementing the

¹ Coast et al. (2008), for instance, assigned relative weights to the dimensions of the ICECAP capability index, in line with the opinions of the older people concerned. Decancq and Michiels (2019) proposed an alternative preference-based method to make comparisons of successful aging using a life satisfaction regression.

factorial survey method in other contexts were optimistic (Sauer et al., 2011; Teti et al., 2016), it remains unclear whether their findings can be generalized to our setting of well-being measurement, to more complex designs (e.g., including vignettes with more levels and dimensions) and to other survey administration modes (e.g., web-based formats).

We will illustrate the potential of the factorial survey method with data from an online longitudinal survey administered by the survey agency Qualtrics. The data was collected among respondents aged 50 years or older in Flanders (the Flemish region of Belgium). Given that respondents were followed over time from May to December 2020, this data contains unique longitudinal information with which we assess the applicability of the factorial survey among older people.

This study makes two contributions. First, we illustrate the usefulness of the factorial survey approach to estimate the relative importance of six well-being dimensions according to older people in Flanders. We found that health, income and social relations were the most important well-being dimensions according to our respondents, followed by leisure and engaging activities. Religion or spirituality appeared to be rather unimportant. Second, we investigate the reliability of these estimates in two different ways.

First, we exploit the longitudinal nature of our data set to perform several within-sample test-retests to assess the consistency of the estimated relative importance weights over time (i.e., temporal reliability) by testing whether the respondents provide consistent answers when completing a similar factorial survey experiment after several weeks. The estimated relative importance weights were consistent over time in 6 out of 10 test-retest analyses performed. We then address the question of how older respondents cope with the complex nature of the factorial survey, by studying two indicators of cognitive load: response time and response consistency.

Response time refers to the time needed for a respondent to address a particular vignette, while response consistency tells us whether a respondent treats each well-being dimension in the same way across all vignette evaluations. For instance, a respondent who treats health as an important determinant of well-being in the first vignette(s), but ignores it when evaluating other vignettes, is said to exhibit inconsistent response behavior. For both indicators, we investigate whether they are systematically related to respondent characteristics and whether they change across the course of vignette evaluations. As respondents progress through the sequence of vignette

ratings, their responses are generally expected to become more precise (i.e., learning effect) until fatigue sets in and precision declines (i.e., fatigue effect) (Auspurg et al., 2009; Sauer et al., 2011).

Overall, we found that the variation in response time and response consistency between respondents with different characteristics was rather small. However, older respondents seemed to have a somewhat lower response speed and the vignette ratings of low income-earners were somewhat less consistent. Moreover, we did not observe any signs of cognitive overload across the sequence of vignette evaluations. On the contrary, the results pointed more in the direction of learning effects. As such, our results are reassuring in supporting the applicability of a factorial survey among older people.

The remainder of this paper is structured as follows. Section 2 will discuss the research methodology, including the design of the factorial survey, the data-collection, the survey sample and the analysis techniques. The results will be presented in Section 3. In Section 4, the findings of this study will be discussed together with its main limitations, while Section 5 will present our conclusions.

METHOD

Design of the factorial survey

The factorial survey is an experimental method that presents respondents with several hypothetical descriptions of an object or situation in order to assess how people make judgments across multidimensional phenomena (Auspurg & Hinz, 2015). In this study, a factorial survey was used to explore the judgments of older people about well-being. Central to this approach is the use of vignettes. A vignette typically contains a combination of randomly selected values (levels) from different dimensions that are assumed to be relevant to the judgment being studied (Auspurg & Hinz, 2015; Rossi & Anderson, 1982). The respondents' task is to express their evaluation of each vignette on a rating scale. During this process, respondents might take multiple dimensions into account and give more weight to the outcomes in some dimensions than others. Using multivariate analysis techniques, researchers can then examine the impact or the relative importance of each level of a dimension on the variation in vignette ratings (Jasso, 2006).

A crucial step in the design of any factorial survey is the selection of dimensions and levels within the vignettes. In this study, an extensive literature review was used to select the life dimensions. More precisely, we included the dimensions "health," "social relations," "income," "leisure," "engagement" and "religion or spirituality" as qualitative studies suggest that these are of major importance in lay views of well-being (see, e.g., Brown et al., 2004; Hung et al., 2010; van Leeuwen et al., 2019). Four levels were specified for each dimension. An overview of the dimensions and levels can be found in Table 1. To familiarize respondents with the dimensions and levels, they were first asked to indicate their own performance on each of the vignette dimensions. Afterwards, participants were presented with the vignettes and asked to indicate, on a 11-point satisfaction scale, how much well-being each hypothetical life situation would bring about according to them. An example can be found in the Appendix.

In a factorial survey with six dimensions and four levels, there are 4,096 possible combinations of dimension levels, which constitute the vignette population. As it is undesirable to completely administer the entire vignette population in the survey (see, e.g., Sauer et al., 2011, on cognitive overload), researchers usually draw a smaller subset of vignettes (Atzmüller & Steiner, 2010). We selected 50 different subsets of 7 vignettes each. The subsets were selected using a computer

algorithm (provided by the SAS macro %Mktex) to create a D-efficient design.² Respondents were randomly assigned to one of the vignette sets.

Table 1. Vignette dimensions and levels

Dimension	Description	Level
Health	Physical or mental health problems	Severe/moderately severe/non-severe/no
Social relations	Contact with family or friends	No/less than once per week/once per week/several times per week
Income	Total net household income per month	€1,500.00/€2,700.00/€3,900.00/€5,000.00 ¹
Leisure	Hobby or leisure activities	No/less than once per week/once per week/several times per week
Engagement	Useful or meaningful activities	No/less than once per week/once per week/several times per week
Religion or spirituality	Time spent on religion or spirituality	No/less than once per week/once per week/several times per week

¹ The income cutoffs were derived from the quintile values of the income distribution of people aged 50 years or older in Belgium (data source: EU-SILC 2018).

² This approach ensures that all effects of the vignette dimensions can be estimated with the maximum amount of statistical precision. It looks, in particular, for a sample which is as close as possible to being balanced (i.e., the levels of each dimension occur equally) and orthogonal (i.e., equal occurrence of each possible combination of levels) (for more details, see Dülmer, 2007). The final D-efficiency value of our sample was 99.99%, indicating an almost perfectly balanced and orthogonal design.

Data collection and survey sample

The factorial survey was implemented in a longitudinal online survey among people aged 50 years or older in Flanders, the Dutch-speaking part of Belgium. Respondents were followed at five different timepoints between May and December 2020. There was an interval of one month between each follow-up survey, with the exception of the last, which took place 10 weeks after wave 4 (see, Table 2, for an overview).

Table 2. Overview waves

	Period	Sample size	
		Total	Recontacts
Baseline	07/05 to 13/05	800	/
Wave 2	10/06 to 21/06	781	452
Wave 3	22/07 to 04/08	827	298
Wave 4	04/09 to 18/09	762	215
Wave 5	01/12 to 17/12	764	154

Participants were recruited from an online panel administered by Qualtrics, a survey agency which employs non-probability sampling strategies in developing its sample frame (Heen et al., 2014). A total of 1,003 respondents participated in the baseline survey (i.e., Wave 1), for which cross-quotas were set on age and gender in order to obtain a balanced sample. All cases in which respondents had obviously been distracted while answering the factorial survey or paused the survey and returned to it at a later point in time were removed. This meant that interviews which took longer than 24 hours to complete were discarded, as well as observations that deviated by twice the standard deviation from the mean reaction time of a single vignette (for the recommendation of this procedure, see Mayerl & Urban, 2008).

As a result, the final sample for the baseline interview comprised 800 respondents. Of these respondents, 19% ($n = 154$) also participated in the four follow-up surveys (i.e., Waves 2 to 5) in which the samples were drawn on a natural fallout basis (i.e., without demographic quotas). More specifically, invitations were first sent to respondents who had previously participated and only when this pool of recontacted respondents was exhausted were new panel members approached. Respondents received an incentive from Qualtrics (either a flat fee or a points system) based on the length of the survey, their specific profile and target acquisition difficulty.

A drop-out analysis showed no significant differences between respondents who participated in all five waves and those who did not (see Supplementary Table 1 in Appendix). We therefore only provide an overview of the sample characteristics from the baseline survey (see Table 3). The average age in the sample was 64.66 (sd = 7.61). Consistent with the average age, the majority of the participants were retired. Approximately half were male, half received higher education, and almost half had incomes of €2,000 or more. As could be expected, representation of lower educated respondents and older adults with a migration background was rather low. Around one third of respondents reported having long-term health problems and another third felt limited in daily activities due to their health.

Table 3. Sample characteristics from the baseline survey (Wave 1)

	%
Male	53.2
Age	
50 to 64 years	47.7
65 to 74 years	41.4
75 years or more	10.9
Highest educational degree	
No or primary	5.1
Secondary	43.4
Higher	51.5
Retired	63.7
Eq. disposable household income	
< €1,500.00	19.3
€1,500.00–€1,999.99	36.4
€2,000.00–€2,999.99	29.5
≥ €3,000.00	14.9
Migration background	7.9
Having long-term health problems	33.9
Being disabled	30.1
Observations	800

Analysis

In the first part of the analysis, we explored the relative importance weights of the well-being dimensions. A factorial survey produces multilevel data, as vignettes are nested within respondents (Hox et al., 1991; Jasso, 2006). Therefore, a multilevel random intercept model was estimated with vignettes as the Level 1 unit of analysis and respondents as Level 2. The estimated model can be specified as follows:

$$S_{ij} = \alpha_0 + \beta x_{ij} + \gamma_{ij} z_{ij} + u_{0j} + e_{0ij}$$

where S_{ij} represents the satisfaction score given by respondent i to vignette j ; x_{ij} is a vector of variables related to the vignette dimension levels (with the worst-off level as the reference case); z_{ij} is a vector of control variables for the vignette position³ and vignette set; u_{0j} is the error component on Level 2 capturing the between-respondents variation; and e_{0ij} is the Level 1 error component measuring the variation within-respondents. Both error terms were assumed to be independently and normally distributed with zero means and constant variances (Hox et al., 1991). The importance weights of the dimension levels can be derived from the estimated β coefficients (Jasso, 2006).

Next, the temporal reliability of the factorial survey was examined by comparing the estimated importance weights between a test and retest period. To do this, we first pooled the data of the test and retest periods (e.g., Waves 1 and 2). We then let the importance weights of the dimension levels interact with a dummy that equaled 0 for the test and 1 for the retest period. Differences in the results of the two time points were formally tested using a Wald test of joint significance of the interaction terms. If the test revealed that there are no significant differences in the importance weights, it would imply that the results of the factorial survey were reliable. In total, we performed 10 different test-retest analyses comparing each possible pair of waves. In order to have a stable sample size across these different test-retests, only respondents who participated in all five waves were included in the analysis.

³ Vignette position was included as a metric variable; dummy coding yields the same results.

Finally, two indicators of cognitive load were computed: response time and response consistency. Response time was available for every single vignette and measured in seconds. Following Sauer et al. (2011), response consistency was obtained from the unexplained variance in the vignette evaluations of each respondent. More specifically, we took the square of the Level 1 residuals from the multilevel random intercept model described above. The higher the consistency of responses, the smaller the amount of variance that was unexplained by the vignette dimensions. Low residual values (approaching zero) thus reflected high levels of response consistency, whereas high residuals indicated low consistency in responses (Sauer et al., 2011).

For both response time and response consistency, we analyzed whether they were systematically related to respondent characteristics and whether they changed across the course of the vignette evaluations. This latter analysis provided insights into the presence of potential learning or fatigue effects across the sequence of vignette evaluations. The following respondent characteristics were taken into account: age (i.e., 50 to 64, 65 to 74, and 75 or more years), education (i.e., none or primary, secondary and higher education), immigration background, equivalized disposable income (i.e., < €1,500.00; €1,500.00–€1,999.99; €2,000.00–€2,999.99; and \geq €3,000.00), being retired, having long-term health problems and being disabled. Income was measured at the household level. The equivalized income was obtained by dividing the household income by the square root of the household size. Migration background was operationalized by country of birth. A person was classified as having a migration background if they or at least one of their parents were born abroad. Having a chronic disease was assessed using a question on whether respondents suffered from physical or mental health problems which had lasted (or were expected to last) six months or more. Being disabled was defined as facing limitations in daily activities due to physical or mental health problems.

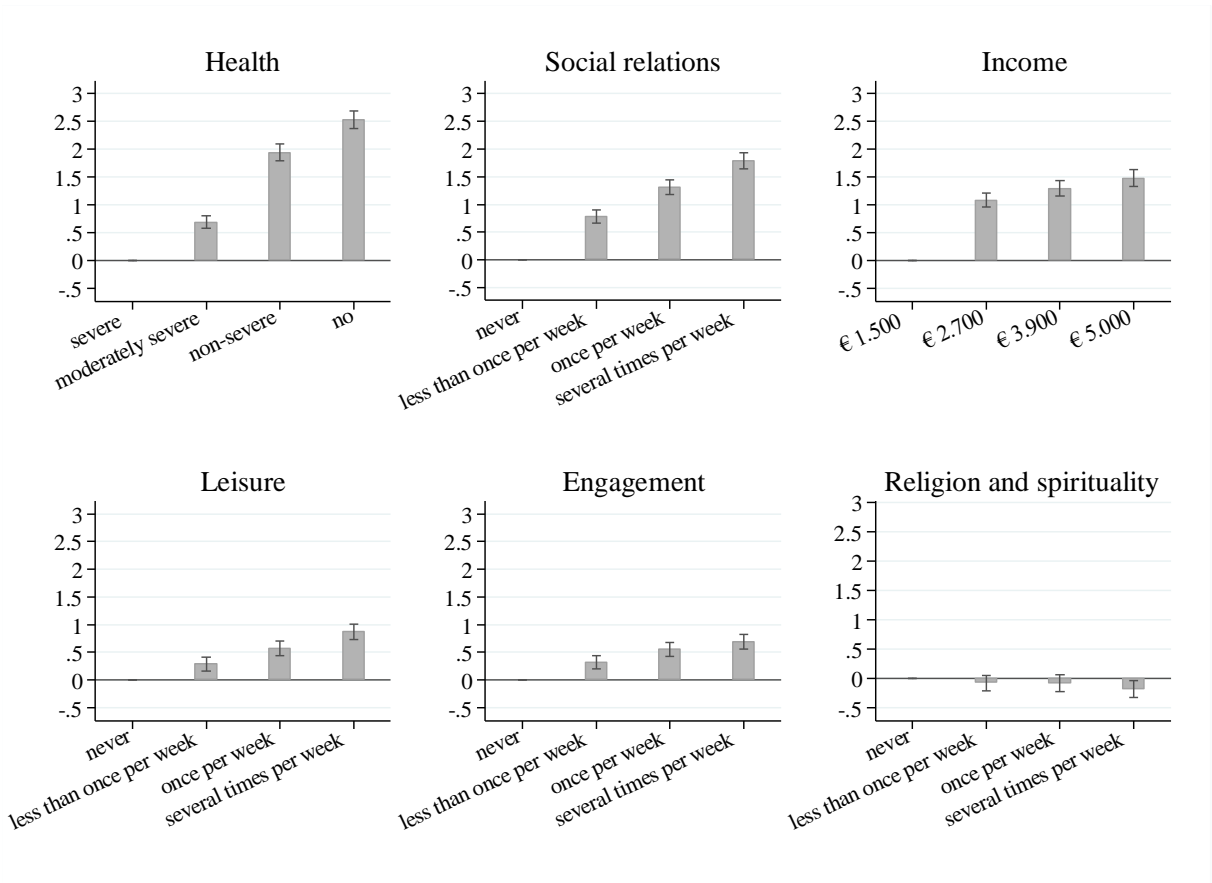
All models were estimated within STATA 16 using the Generalized Least Squares (GLS) estimator with cluster-robust standard errors.

RESULTS

Relative importance of well-being dimensions

A multilevel regression analysis of vignette evaluations on the level of the dimensions was performed to examine the potential of the factorial survey to elicit the view of older people on well-being. Regression coefficients are particularly informative in this regard as they reveal the net increase or decrease in the overall vignette score of a particular dimension level compared to the lowest level (i.e., reference category). In other words, these coefficients represent the relative importance weights of a particular dimension level. The regression coefficients along with the standard errors can be found in the Appendix (see Supplementary Table 2 in the Appendix).

Figure 1. Visualization of the relative importance weights (and 95% confidence interval) of the dimension levels (n = 800)



Note: Based on a multilevel regression (GLS) with robust standard errors. Relative importance weights were estimated including controls for design effects (i.e., vignette position and dummy variables for vignette set).

With the exception of religion and spirituality, all importance weights were significantly different from zero and pointed in the anticipated direction. Indeed, Figure 1 shows that respondents attribute higher weights to more beneficial levels. The results for income, however, indicate that respondents clearly dislike the lowest level, but the additional gain of more favorable levels is relatively small compared to changes within other dimensions.

The magnitude of the importance weights not only varied within dimensions but also across dimensions. The highest importance was attributed to being perfectly healthy. A change from severe health problems to no health problems would, according to our participants, increase their well-being by 2.53 points (on the 11-point satisfaction scale). Respondents also assigned a high weight to changes from the worst to the best level within the dimensions related to social relations and income (1.79 and 1.48 points respectively). Overall, however, the perceived importance of these dimensions was somewhat lower than for health. For example, going from severe health problems to perfect health yielded approximately 1.4 ($2.53/1.79$) times as much well-being as increasing the frequency of social contacts from never to several times per week. Likewise, a change from severe health problems to perfect health resulted in 1.7 ($2.53/1.48$) times as much well-being as an income increase of €3,500 per month.

Leisure and engagement were perceived to be almost equally important, but their importance weights were again markedly lower than those of health, social relations and income. Investing time in religion or spirituality appears to generate little well-being according to our respondents, except when the frequency increases to several times per week. At that point, the weight becomes significant and the vignette rating drops by 0.18 on the 11-point satisfaction scale.

Temporal reliability

Thus far, we have explored the relative importance weights of the well-being dimensions. In order to examine whether the estimated importance weights were reliable, we estimated a pooled model with additional interaction terms between the importance weights and a dummy that equaled 0 for the test period and 1 for the retest period. Subsequently, a Wald test of joint significance of the interactions was used to evaluate whether the results from the test and retest diverged (see Table 4).

Table 4. Wald test of equal importance weights between test and retest (n = 154)

		Test			
		Wave 1	Wave 2	Wave 3	Wave 4
Retest	Wave 2	** chi ² = 37.48			
	Wave 3	** chi ² = 38.32	ns chi ² = 24.31		
	Wave 4	* chi ² = 31.68	ns chi ² = 20.01	ns chi ² = 20.69	
	Wave 5	ns chi ² = 22.21	ns chi ² = 22.50	ns chi ² = 18.73	* chi ² = 29.14

Note: df = 18; ns $p > 0.05$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Only respondents who participated in all five waves were included in the analysis (n = 154).⁴

In 6 out of 10 cases, the null hypothesis of similar parameter estimates could not be rejected at the 5% significance level – providing support for the temporal reliability of the importance weights. However, the importance weights expressed in Wave 1 proved to be statistically different from those in Waves 2, 3 and 4. In addition, we found that the importance weights changed significantly between Wave 4 and Wave 5. The most noticeable difference over time was related to the income dimension. More precisely, we found that income was perceived to be less important in Wave 1 than in the consecutive waves (results not reported here).

Feasibility

⁴ Using the largest possible sample for each pair of waves provided similar results – except that there was no significant difference found between Waves 1 and 2 (see Supplementary Table 3 in the Appendix).

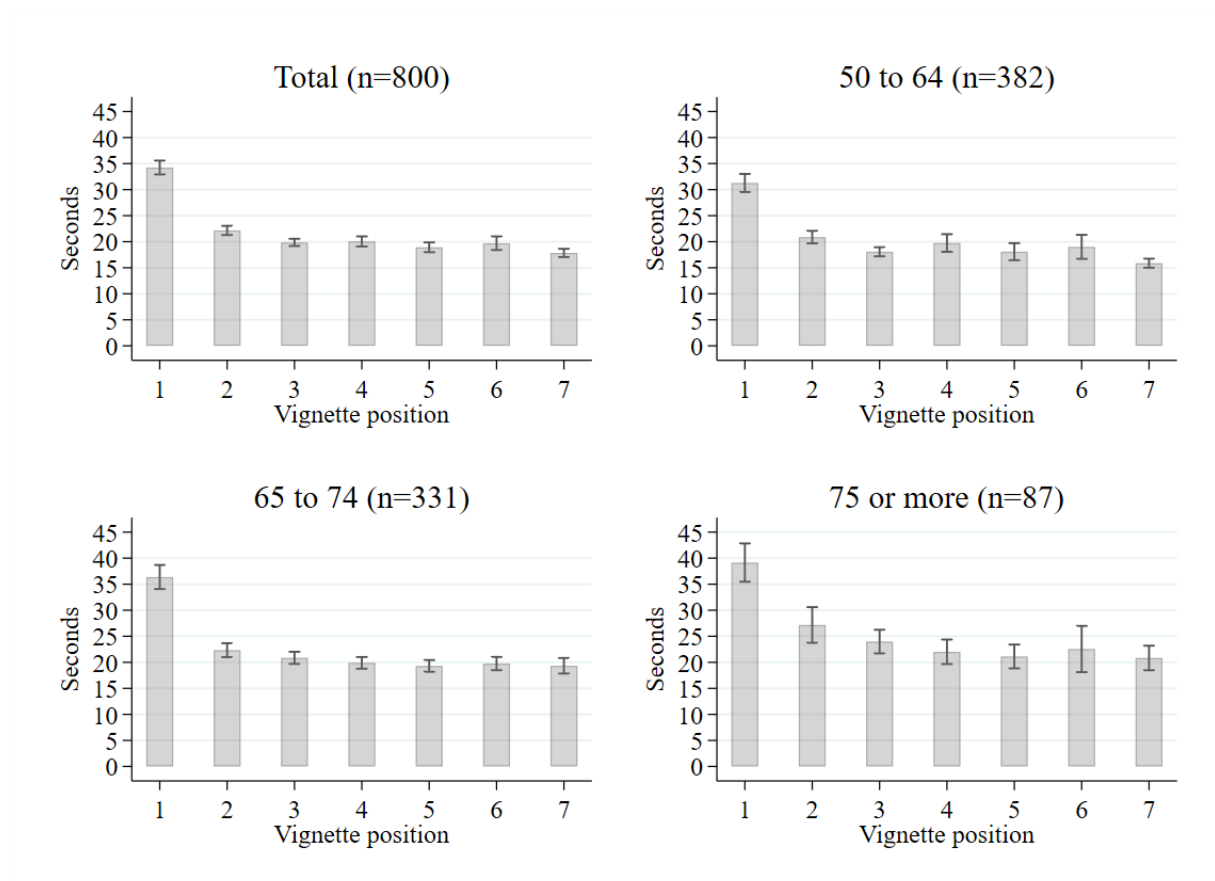
Response behavior provides valuable information on the cognitive load of the factorial survey and allows us to detect potential problems in handling the vignette evaluations. Below, we take a closer look at two indicators of response behavior: response time and response consistency.

Response time

Overall, respondents needed 2.5 minutes to complete the entire vignette module. This is equivalent to an average of 21.8 seconds per vignette. Model 1 in Table 5 displays the effects of respondent characteristics on response time per vignette. The results suggest that older respondents required more time than younger respondents. More precisely, it took the oldest participants (i.e., 75 years or older) almost 5 seconds longer to rate a single vignette compared to their youngest counterparts (i.e., 50 to 64 years). Likewise, the average response time per vignette was somewhat longer among respondents aged 65 to 74 years than among those aged 50 to 64 years. Moreover, we found that women and highly educated respondents took more time to rate the vignettes, compared to men and lower educated respondents. However, in contrast to the effect of age, these findings were not robust across waves (see Supplementary Table 4 in the Appendix).

Figure 2 shows the evolution of response time across the sequence of vignettes. Given that older participants needed more time to evaluate a vignette, the age variable warrants special consideration. In general, we observed that the response speed was lowest at the beginning of the vignette module. Afterwards, the response time decreased rapidly and stabilized after the third vignette. Across the entire sequence of vignettes, the response speed of older respondents was generally lower than the response speed of younger respondents. However, the same pattern, that is, a sharp decline of response time within the first part of the vignette module, was observed in all age groups.

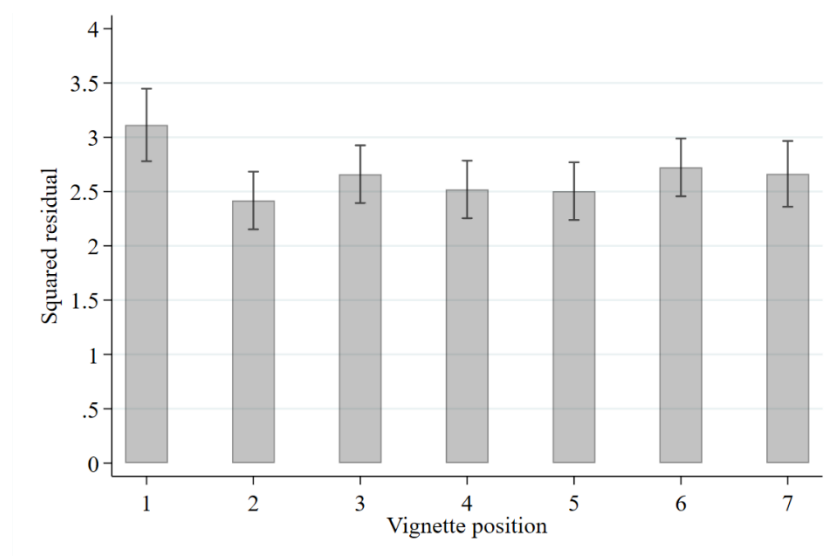
Figure 2. Average response time (and 95% confidence intervals) by vignette position



Response consistency

Model 2 in Table 5 predicts the absolute value of the squared residual, which was used as an indicator of response consistency by respondent characteristics. A positive coefficient reflects an increase in the inconsistency of the vignette ratings. Although older age groups showed a longer response time, there was no indication that their responses were less consistent. Significant effects did emerge, however, for income and the variable related to having a chronic disease. More specifically, those with a higher income (i.e., between €2,000.00–€2,999.99 or ≥ €3,000.00) were more consistent in evaluating the vignettes than individuals with a low income (i.e., < €1,500.00). The responses of participants with long-term mental or physical health problems, on the other hand, were less consistent than those of their more fortunate counterparts. The latter effect was not confirmed in other waves (see Supplementary Table 5 in the Appendix).

Figure 3. Average response consistency (and 95% confidence intervals) by vignette position (n = 800)



As can be seen from Figure 3, the level of response inconsistency drops significantly after the first vignette, but remains relatively stable afterwards. After the first vignette, respondents are thus able to make efficient judgments at a stable level of response consistency. This pattern was observed among all respondents – including the oldest, lowest educated and those with physical or mental impairments (not reported here). At first glance, these results imply the absence of fatigue effects and hint in the direction of learning effects.⁵

⁵ Following Sauer et al. (2014), we tested whether the observed learning effects were not the result of simplified decision heuristics (such as fading out of some dimensions) by estimating separate regressions for different parts of the vignette module (i.e., the first versus the last three vignette ratings). The results are presented in the Appendix (see Supplementary Table 6). The number of significant estimates, which served as a proxy for the amount of information taken into account by the respondent, as well as their effect sizes, remained stable over the sequence of vignette evaluations. These results suggest that respondents did not apply any decision heuristic to simplify their answers as they progressed through our factorial survey experiment.

Table 5. Random intercept models of response time (RT) and response consistency (RC)

	Model 1		Model 2	
	RT		RC	
	b	Se	b	Se
Male (1 = yes)	-1.503*	(0.670)	0.167	(0.156)
Age (ref. 50 to 64 years)				
65 to 74 years	2.014*	(0.831)	-0.156	(0.229)
75 years and older	4.912***	(1.290)	0.174	(0.365)
Education (ref. no or primary)				
Secondary	2.136	(1.101)	0.096	(0.370)
Higher	2.465*	(1.089)	-0.161	(0.364)
Retired (1 = yes)	0.183	(0.877)	0.036	(0.243)
Eq. disposable household income (ref. < €1,500.00)				
€1,500.00–€1,999.99	-0.508	(0.826)	-0.069	(0.234)
€2,000.00–€2,999.99	0.235	(0.974)	-0.495*	(0.229)
≥ €3,000.00	-1.472	(1.144)	-0.621*	(0.242)
Migration background (1 = yes)	-1.943	(1.146)	0.498	(0.390)
Having long-term health problems (1 = yes)	-0.703	(0.933)	0.573*	(0.234)
Being disabled (1 = yes)	1.877	0.984	0.102	0.234
Baseline speed ¹	0.000	0.000		
Constant	24.270***	(1.831)	2.057***	(0.603)
Sigma_u	7.486		1.553	
Sigma_e	12.764		3.727	
Wald chi ²	811.002		117.012	
p-value	0.000		0.000	
R ²	11.6%		3.3%	
Respondents	800		800	
Vignettes	5,600		5,600	

Note: Based on a multilevel regression (GLS) with robust standard errors.

Tested with controls for design effects (i.e., vignette position and dummy variables for vignette set).

¹ Baseline speed is defined as the time that a person needs to answer questions, independent of the content. It was measured by subtracting the response time of the vignette module from the entire survey length.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

DISCUSSION

In this study, we investigated the potential of a factorial survey to estimate the relative importance of well-being dimensions among older people. Overall, the results confirmed earlier findings that, according to the older population, well-being is multidimensional (Bowling & Dieppe, 2005; Cosco et al., 2013; Hung et al., 2010; Jopp et al., 2015; Phelan et al., 2004). Indeed, our results suggest that health, social relations, financial resources, leisure time and active engagement are all important to the well-being of older people. Religion and spirituality, however, appeared to be less important in their view on well-being. One plausible explanation for this may be found in the process of secularization that has marked Western European societies, including Flanders, since the 1960s (Dobbelaere, 2002). As such, the factorial survey demonstrated its capacity to distinguish important from unimportant aspects. As the example of religion and spirituality shows, the latter may potentially be influenced by factors such as the surrounding culture.

In addition, the factorial survey revealed a certain hierarchy in the attributed importance of well-being dimensions. It is clear from our results that not all dimensions are equally important: health generally proved to be most important, followed by social contacts and income. Leisure time and engaging activities, by contrast, played a somewhat smaller role. Interestingly, the results also illustrated that the relative importance of a well-being dimension strongly depends on its particular level. In line with the economic literature on the marginal utility of income (see, e.g., Layard et al., 2008), we found that respondents clearly disliked being poor (i.e., the lowest level), but the additional gains of higher income levels were relatively small. By measuring the view of older people on well-being based on concrete outcome levels in the different well-being dimensions, the factorial survey allowed us to disentangle the complex trade-offs between dimensions. As a result, the derived importance weights were, on the one hand, sensitive to the way the dimension levels were operationalized, but provided more precise information about the implied trade-offs made by respondents than rating or ranking general dimensions as single items.

The estimated importance weights of the life dimensions were not only plausible, but also reliable. In 6 out of 10 test-retest analyses performed, the importance weights proved to be consistent over time. In fact, the level of consistency between the tests and retests was quite

remarkable considering that respondents evaluated a different set of vignettes each time. In the other four cases, the results of the tests and retests were slightly different, mainly because the attributed importance of income was somewhat lower in Wave 1 compared to the other waves. Further research is needed to interpret this result conclusively. However, due to the COVID-19 pandemic and subsequent policy measures to control the spread of the coronavirus, it is likely that people have adapted their life goals to reflect changing life circumstances (for a discussion, see also Bland, 2020). Given that a large part of our sample was retired, and pensions remained stable, income was one of the few life aspects that did not dramatically change during the COVID-19 crisis in Flanders. Perhaps this may explain why income was perceived as less important in Wave 1.

The analyses of response time and response consistency provided further evidence that older respondents coped well with the complexity of the factorial survey. Overall, variation in response time between respondents with different characteristics was small, although there was a tendency to lower response speed in the older age groups. According to Auspurg et al. (2009) and Sauer et al. (2011), such an age effect is inherent to any question type and therefore not indicative of problems specific to a factorial survey. In line with a previous feasibility study (Teti et al., 2016), respondents from different age groups and educational backgrounds showed similar levels of response consistency. Regarding household income, we did find that the inconsistency in responses was somewhat higher among low-income earners, again confirming the results of Teti et al. (2016).

Finally, we found no signs of cognitive overload across the sequence of vignette evaluations. On the contrary, the results pointed more in the direction of learning effects and were thus reassuring for the applicability of a factorial survey among older people. For both response time and response inconsistency, a substantive drop was observed after the first vignette: obviously respondents needed some time to become familiar with the rating task at hand. Nevertheless, the first vignette evaluations were already acceptable in terms of response time and consistency. Two remarks need to be made in this respect. First, it is important to emphasize that at the beginning of the survey respondents were asked to indicate their own level in each vignette dimension and to rate a vignette of their own life. Respondents were thus already familiar with the vignette descriptions before the actual factorial survey experiment started. Second, respondents were presented with only seven different vignettes. It could thus be true that

cognitive overstrain may occur if respondents have to evaluate a higher number of vignettes (Auspurg et al., 2009; Sauer et al., 2011; Teti et al., 2016).

The results of this study should be interpreted in light of several limitations. First of all, our study was conducted in an online setting, in which respondents were drawn from a non-probability panel. As certain subpopulations may self-select into such panels, the generalizability of our results to the general population of older people might potentially be affected. In fact, our sample was predominantly white, highly educated and in good health. Moreover, it is probable that the Quattrics panel consists mainly of experienced survey participants, who are likely to be more familiar with cognitively demanding survey questions than unexperienced individuals would be. Future research is needed to investigate whether our findings can be replicated among more heterogeneous and representative samples of the older population, and in different geographic regions.

CONCLUSION

Due to the rapid ageing of our society, the need to evaluate health and social care services for older people is expected to grow considerably. An accurate measurement of well-being, including the weighing of well-being dimensions, is indispensable in this regard, and choosing the appropriate methodology to do so has become all the more relevant (Himmeler et al., 2021). Against this background, this study investigated the potential of a factorial survey to derive the relative importance of well-being dimensions among older people. Overall, the estimated importance of the dimensions proved to be plausible and reliable. In addition, we found that the target population of older people was able to cope with the complexity of the factorial survey and produced a high level of response consistency within an acceptable amount of time. We believe, therefore, that factorial surveys offer us a promising way forward in eliciting the views of older people on well-being at an older age and, hence, in developing policies that matter to them.

REFERENCES

- 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. doi:10.1027/1614-2241/a000014
- Auspurg, K., & Hinz, T. (2015). *Quantitative Applications in the Social Sciences: Factorial survey experiments*. Thousand Oaks, CA: SAGE Publications.
- Auspurg, K., Hinz, T., & Liebig, S. (2009). Complexity, Learning Effects and Plausibility of Vignettes in the Factorial Survey Design. *methods, data, analyses*, 3(1), 59-96. doi:10.12758/mda.2009.003
- Bland, A. M. (2020). Existential Givens in the COVID-19 Crisis. *Journal of Humanistic Psychology*, 60(5), 710–724. doi:10.1177/0022167820940186
- Bowling, A., & Dieppe, P. (2005). What is successful ageing and who should define it? *BMJ (Clinical Research ed.)*, 331(7531), 1548–1551. doi:10.1136/bmj.331.7531.1548
- Bowling, A., & Gabriel, Z. (2004). An Integrational Model of Quality of Life in Older Age. Results from the ESRC/MRC HSRC Quality of Life Survey in Britain. *Social Indicators Research*, 69(1), 1–36. doi: 10.1023/B:SOCI.0000032656.01524.07
- Brown, J., Bowling, A., & Flynn, T. (2004). Models of quality of life: A taxonomy, overview and systematic review of the literature. Report commissioned by European Forum on Population Ageing Research. Sheffield: University of Sheffield. Retrieved from <http://www.shef.ac.uk/ageingresearch>
- Coast, J., Flynn, T. N., Natarajan, L., Sproston, K., Lewis, J., Louviere, J. J., & Peters, T. J. (2008). Valuing the ICECAP capability index for older people. *Social Science & Medicine*, 67(5), 874–882. doi:10.1016/j.socscimed.2008.05.015
- Cosco, T. D., Prina, A. M., Perales, J., Stephan, B. C., & Brayne, C. (2013). Lay perspectives of successful ageing: a systematic review and meta-ethnography. *BMJ Open*, 3. doi:10.1136/bmjopen-2013-002710

- Cosco, T. D., Prina, A. M., Perales, J., Stephan, B. C., & Brayne, C. (2014). Operational Definitions of Successful Aging: a Systematic Review. *International Psychogeriatrics*, 26(3), 373–381. doi:10.1017/S1041610213002287
- Decancq, K., & Michiels, A. (2019). Measuring Successful Aging With Respect for Preferences of Older People. *The journals of Gerontology. Series B, Psychological sciences and social sciences*, 74(2), 364–372. doi:10.1093/geronb/gbx060
- Decade of healthy ageing: baseline report. Geneva: World Health Organization; 2020. Licence: CC BY-NC-SA 3.0 IGO.
- Depp, C. A., & Jeste, D. V. (2006). Definitions and Predictors of Successful Aging: A Comprehensive Review of Larger Quantitative Studies. *The American Journal of Geriatric Psychiatry*, 14(1), 6–20. doi:10.1097/01.JGP.0000192501.03069.bc
- Dobbelaere, K. (2002). *Secularization: An Analysis at Three Levels*. Peter Lang.
- Dülmer, H. (2007). Experimental Plans in Factorial Surveys: Random or Quota Design? *Sociological Methods & Research*, 35(3), 382–409. doi:10.1177/0049124106292367
- Fernández-Ballesteros, R. (2011). Quality of Life in Old Age: Problematic Issues. *Applied Research in Quality of Life*, 6(1), 21-40. doi:10.1007/s11482-010-9110-x
- Gabriel, Z. & Bowling, A. (2004). Quality of Life from the Perspectives of Older People. *Ageing and Society*, 24(5). 675–691. doi:10.1017/S0144686X03001582
- Hackert, M.Q.N, Brouwer, W.B.F, Hoefman, R.J, & van Exel, N.J.A. (2019). Views of older people in the Netherlands on wellbeing: A Q-methodology study. *Social Science & Medicine*, 240. doi:10.1016/j.socscimed.2019.112535
- Heen, M.S.J., Lieberman, J. D., & Miethe, T. D. (2014). A Comparison of Different Online Sampling Approaches for Generating National Samples (Report No. CCJP 2014–01). Las Vegas: University of Nevada, Las Vegas, Center for Crime and Justice Policy. Retrieved from

http://www.unlv.edu/sites/default/files/page_files/27/ComparisonDifferentOnlineSampling.pdf

- Himmler, S., Soekhai, V., van Exel, J., & Brouwer, W. (2021). What works better for preference elicitation among older people? Cognitive burden of discrete choice experiment and case 2 best-worst scaling in an online setting. *Journal of Choice Modelling*, 38, Article 100265. doi:10.1016/j.jocm.2020.100265
- 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
- Hsieh, C.M. (2005). Age and relative importance of major life domains. *Journal of Aging Studies*, 19(4), 503-5012. doi.org/10.1016/j.jaging.2005.07.001.
- Hung, L., Kempen, G., & De Vries, N. (2010). Cross-cultural comparison between academic and lay views of healthy ageing: A literature review. *Ageing and Society*, 30(8), 1373-1391. doi:10.1017/S0144686X10000589
- Jasso, G. (2006). Factorial Survey Methods for Studying Beliefs and Judgments. *Sociological Methods & Research*, 34(3), 334–423. doi:10.1177/0049124105283121
- Jopp, D. S., Wozniak, D., Damarin, A. K., De Feo, M., Jung, S., & Jeswani, S. (2015). How Could Lay Perspectives on Successful Aging Complement Scientific Theory? Findings From a U.S. and a German Life-Span Sample. *The Gerontologist*, 55(1), 91–106. doi:10.1093/geront/gnu059
- Layard, R., Mayraz, G., & Nickell, S. (2008). The marginal utility of income. *Journal of Public Economics*, 92(8-9), 1846-1857. doi.org/10.1016/j.jpubeco.2008.01.007
- Mayerl, J., & Urban, D. (2008). Antwortreaktionszeiten in Survey-Analysen: Messung, Auswertung und Anwendungen. Wiesbaden:VS Verlag für Sozialwissenschaften.
- Phelan, E. A., Anderson, L. A., LaCroix, A. Z., & Larson, E. B. (2004). Older Adults' Views of "Successful Aging" – How Do They Compare with Researchers' Definitions? *Journal*

- of the American Geriatrics Society*, 52(2), 211–216. doi:10.1111/j.1532-5415.2004.52056.x
- Phelan, E. A., & Larson, E. B. (2002). “Successful aging”—where next? *Journal of the American Geriatrics Society*, 50(7), 1306–1308. doi:10.1046/j.1532-5415.2002.50324.x
- Pruchno, R. A. (2015). Successful Aging: Contentious Past, Productive Future. *The Gerontologist*, 55(1), 1–4. doi:10.1093/geront/gnv002
- Pruchno, R. A., Wilson-Genderson, M., & Cartwright, F. (2010). A Two-Factor Model of Successful Aging. *The Journals of Gerontology. Series B, Psychological sciences and social sciences*, 65(6), 671–679. doi:10.1093/geronb/gbq051
- 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.
- Rowe, J. W., & Kahn, R. L. (1997). Successful Aging. *The Gerontologist*, 37(4), 433–440. doi.org/10.1093/geront/37.4.433
- 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. doi:10.18148/srm/2011.v5i3.4625
- Sauer, C., Auspurg, K., Hinz, T., Liebig, S., & Schupp, J. (2014). Methods effects in factorial surveys: an analysis of respondents’ comments, interviewers’ assessments, and response behavior. SOEPPaper No. 629, 31. doi: 10.2139/ssrn.2399404
- Strawbridge, W. J., Wallhagen, M. I., & Cohen, R. D. (2002). Successful Aging and Well-Being: Self-Rated Compared With Rowe and Kahn. *The Gerontologist*, 42(6), 727–733. doi:10.1093/geront/42.6.727
- 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. doi:10.1177/0164027515600767

- United Nations Department of Economic and Social Affairs, Population Division (2020). *World Population Ageing 2020 Highlights: Living arrangements of older people*. New York; United Nations (ST/ESA/SER.A/451).
- van Leeuwen, K. M., van Loon, M. S., van Nes, F. A., Bosmans, J. E., de Vet, H., Ket, J., Widdershoven, G., & Ostelo, R. (2019). What does quality of life mean to older adults? A thematic synthesis. *PloS one*, *14*(3), e0213263. doi:10.1371/journal.pone.0213263
- Wallander, L. (2009). 25 years of factorial surveys in sociology: A review. *Social Science Research*, *38*(3), 505–520. doi:10.1016/j.ssresearch.2009.03.004
- Whitley, E., Benzeval, M., & Popham, F. (2020). Population Priorities for Successful Aging: A Randomized Vignette Experiment. *The journals of Gerontology. Series B, Psychological sciences and social sciences*, *75*(2), 293–302. doi:10.1093/geronb/gby060
- Wilhelmson, K., Andersson, C., Waern, M., & Allebeck, P. (2005). Elderly people's perspectives on quality of life. *Ageing and Society*, *25*(4), 585–600. doi:10.1017/S0144686X05003454



APPENDIX

Figure 1. Example vignette

Please read the following life description carefully.

*“You have [moderately severe] physical or mental health problems.
You have [several times per week] contact with family or friends.
The total net household income is [€5,000.00].
You do [once per week] a hobby or leisure activity.
You do [several times per week] a useful or meaningful activity.
You spend [less than once per week] time on religion or spirituality.”*

How satisfied would you be if you were in this situation?

 0 1 2 3 4 5 6 7 8 9 10 

Note: The words between brackets are the levels that varied experimentally from vignette to vignette.
The order of the dimensions within the vignettes varied across the vignette sets to avoid potential order effects.

Table 1. Logistic regression of dropping out of the survey (n = 800)

	Odds ratio	se
Male (1 = yes)	0.879	(0.160)
Age (ref. 50 to 64 years)		
65 to 74 years	1.235	(0.304)
75 years and older	0.967	(0.320)
Education (ref. no or primary)		
Secondary	1.426	(0.528)
Higher	2.064	(0.785)
Retired (1 = yes)	0.777	(0.199)
Eq. disposable household income (ref. < €1,500.00)		
€1,500.00–€1,999.99	1.361	(0.334)
€2,000.00–€2,999.99	1.308	(0.351)
≥ €3,000.00	1.465	(0.481)
Migration background (1 = yes)	1.331	(0.484)
Having long-term health problems (1 = yes)	1.243	(0.329)
Being disabled (1 = yes)	1.061	(0.279)
Loglikelihood	-2,698.435	
R ²	1.6%	
Respondents	800	
Vignettes	5,600	

Note: Based on a logistic regression with robust standard errors.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2. The relative importance weights of the dimension levels

	b	se
Physical or mental health problems (ref. severe)		
Moderately severe	0.693 ***	(0.060)
Non-severe	1.941 ***	(0.077)
No	2.529 ***	(0.079)
Having contact with family or friends (ref. never)		
< 1 per week	0.782 ***	(0.063)
1 per week	1.313 ***	(0.070)
> 1 time per week	1.791 ***	(0.074)
Household income (ref. €1,500.00)		
€2,700.00	1.089 ***	(0.066)
€3,900.00	1.298 ***	(0.072)
€5,000.00	1.483 ***	(0.077)
Doing hobbies or leisure activities (ref. never)		
< 1 per week	0.292 ***	(0.065)
1 per week	0.568 ***	(0.068)
> 1 time per week	0.872 ***	(0.070)
Doing useful or meaningful activities (ref. never)		
< 1 per week	0.324 ***	(0.062)
1 per week	0.554 ***	(0.065)
> 1 time per week	0.692 ***	(0.065)
Spending time on religion or spirituality (ref. never)		
< 1 per week	-0.078	(0.067)
1 per week	-0.082	(0.073)
> 1 time per week	-0.186 *	(0.072)
Constant	0.252	(0.390)
Sigma_u	1.283	
Sigma_e	1.734	
Wald chi ²	2916.542	
p-value	0.000	
R ²	32.0%	
Respondents	800	
Vignettes	5,600	

Note: Based on a multilevel regression (GLS) with robust standard errors.

Tested with controls for design effects (i.e., vignette position and dummy variables for vignette set).

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3. Wald test of equal importance weights between test and retest (total sample)

		Test			
		Wave 1	Wave 2	Wave 3	Wave 4
Retest	Wave 2 (n = 452)	ns chi ² = 25.57			
	Wave 3 (n = 298)	*** chi ² = 56.00	ns chi ² = 20.35		
	Wave 4 (n = 215)	* chi ² = 30.90	ns chi ² = 18.09	ns chi ² = 15.20	
	Wave 5 (n = 154)	ns chi ² = 22.21	ns chi ² = 22.50	ns chi ² = 18.73	* chi ² = 29.14

Note: df = 18; ^{ns} $p > 0.05$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4. Random intercept models of response time across waves

	Wave 1		Wave 2		Wave 3		Wave 4		Wave 5	
	b	se	b	se	b	se	b	Se	b	se
Male (1 = yes)	-1.503 *	(0.670)	-2.177 **	(0.692)	-0.782	(0.721)	0.564	(0.664)	0.183	(0.675)
Age (ref. 50 to 64 years)										
65 to 74 years	2.014 *	(0.831)	2.638 **	(0.966)	0.759	(0.997)	0.363	(0.942)	1.664	(0.858)
75 years and older	4.912 ***	(1.290)	6.129 ***	(1.523)	3.447 *	(1.505)	2.829 *	(1.217)	3.034 **	(1.060)
Education (ref. no or primary)										
Secondary	2.136	(1.101)	1.484	(1.398)	0.062	(1.492)	0.042	(1.932)	-0.848	(1.747)
Higher	2.465 *	(1.089)	0.840	(1.370)	0.203	(1.576)	-0.184	(1.957)	-0.187	(1.789)
Retired (1 = yes)	0.183	(0.877)	0.293	(1.000)	1.143	(1.107)	1.087	(1.093)	-0.746	(0.927)
Eq. disposable household income (ref. < €1,500.00)										
€1,500.00–€1,999.99	-0.508	(0.826)	1.145	(0.826)	0.309	(0.998)	1.490	(0.974)	-0.148	(1.052)
€2,000.00–€2,999.99	0.235	(0.974)	1.417	(0.962)	0.310	(1.021)	0.658	(1.026)	-0.331	(1.067)
≥ €3,000.00	-1.472	(1.144)	0.120	(1.075)	-1.572	(1.179)	-0.856	(1.057)	-1.802	(1.297)
Migration background (1 = yes)	-1.943	(1.146)	0.764	(1.904)	-1.865	(1.039)	-1.482	(1.769)	0.211	(1.337)
Having long-term health problems (1 = yes)	-0.703	(0.933)	-0.997	(0.814)	-1.472	(0.886)	-1.940 **	(0.735)	1.114	(0.886)
Being disabled (1 = yes)	1.877	(0.984)	2.409 **	(0.912)	2.411 *	(0.943)	2.231 **	(0.851)	0.712	(0.941)
Baseline speed ¹	0.000	(0.000)	0.000 **	(0.000)	0.000	(0.000)	0.000 +	(0.000)	-0.000	(0.000)
Constant	24.270 ***	(1.831)	28.406 ***	(2.779)	27.370 ***	(2.502)	26.402 ***	(2.209)	27.184 ***	(2.778)
Sigma_u	7.486									
Sigma_e	12.764		15.683		17.563		17.192		16.597	
Wald chi ²	811.002		954.446		956.316		704.569		594.333	
p-value	0.000		0.000		0.000		0.000		0.000	
R ²	11.6%		12.8%		0.097		0.078		0.080	
Respondents	800		781		827		761		763	
Vignettes	5,600		5,467		5,789		5,327		5,341	

Note: Based on a multilevel regression (GLS) with robust standard errors. Tested with controls for design effects (i.e., vignette position and dummy variables for vignette set).

¹ Baseline speed was defined as the general mental speed that a person needs to answer questions, independent of the content of the questions. It was measured by subtracting the response time of the vignette module from the entire survey length. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5. Random intercept models of response consistency across waves

	Wave 1		Wave 2		Wave 3		Wave 4		Wave 5	
	b	se	b	se	b	se	b	se	b	se
Male (1 = yes)	0.167	(0.156)	0.269	(0.154)	0.347 *	(0.145)	0.065	(0.157)	0.298	(0.162)
Age (ref. 50 to 64 years)										
65 to 74 years	-0.156	(0.229)	0.291	(0.214)	0.364	(0.204)	-0.062	(0.236)	-0.079	(0.204)
75 years and older	0.174	(0.365)	0.256	(0.335)	0.546	(0.297)	0.062	(0.277)	0.163	(0.301)
Education (ref. no or primary)										
Secondary	0.096	(0.370)	-0.403	(0.487)	-0.301	(0.381)	-0.634	(0.457)	-0.394	(0.401)
Higher	-0.161	(0.364)	-0.908	(0.494)	-0.625	(0.386)	-1.126 *	(0.455)	-0.596	(0.398)
Retired (1 = yes)	0.036	(0.243)	-0.429 *	(0.216)	-0.535 *	(0.212)	-0.033	(0.233)	-0.426	(0.235)
Eq. disposable household income (ref. < €1,500.00)										
€1,500.00–€1,999.99	-0.069	(0.234)	-0.689 **	(0.250)	-0.054	(0.254)	-0.348	(0.274)	-0.579	(0.308)
€2,000.00–€2,999.99	-0.495 *	(0.229)	-0.960 ***	(0.254)	-0.272	(0.256)	-0.917 **	(0.281)	-0.962 **	(0.324)
≥ €3,000.00	-0.621 *	(0.242)	-1.015 ***	(0.281)	-0.219	(0.272)	-0.516	(0.299)	-0.918 **	(0.353)
Migration background (1 = yes)	0.498	(0.390)	-0.317	(0.391)	-0.254	(0.258)	-0.521	(0.303)	-0.357	(0.304)
Having long-term health problems (1 = yes)	0.573 *	(0.234)	-0.027	(0.199)	-0.197	(0.193)	0.342	(0.216)	0.256	(0.187)
Being disabled (1 = yes)	0.102	0.234	0.007	(0.217)	0.454 *	(0.209)	-0.082	(0.210)	0.038	(0.196)
Constant	2.057 ***	(0.603)	4.422 ***	(0.619)	3.318 ***	(0.627)	3.944 ***	(0.695)	4.477 ***	(0.958)
Sigma_u	1.553		1.481		1.443		1.543		1.672	
Sigma_e	3.727		4.013		3.937		3.920		4.003	
Wald chi ²	117.012		131.787		116.621		156.464		118.858	
p-value	0.000		0.000		0.000		0.000		0.000	
R ²	3.3%		3.9%		3.3%		3.8%		3.2%	
Respondents	800		781		827		761		763	
Vignettes	5,600		5,467		5,789		5,327		5,341	

Note: Based on a multilevel regression (GLS) with robust standard errors.

Tested with controls for design effects (i.e., vignette position and dummy variables for vignette set).

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 6. Random intercept model of vignette judgments: 1st vs. 2nd half of the vignette module

	1 st , 2 nd , and 3 rd vignette		5 th , 6 th , and 7 th vignette	
	b	se	b	se
Physical or mental health problems (ref. severe)				
Moderately severe	0.726***	(0.112)	0.540***	(0.106)
Non-severe	1.815***	(0.122)	1.768***	(0.130)
No	2.469***	(0.113)	2.355***	(0.136)
Having contact with family or friends(ref. never)				
< 1 per week	1.034***	(0.129)	0.617***	(0.107)
1 per week	1.335***	(0.134)	1.270***	(0.128)
> 1 time per week	1.662***	(0.131)	1.752***	(0.118)
Household income (ref. €1.500,00)				
€2.700,00	1.018***	(0.123)	1.126***	(0.130)
€3.900,00	1.136***	(0.123)	1.332***	(0.131)
€5.000,00	1.268***	(0.125)	1.666***	(0.140)
Doing hobbies or leisure activities (ref. never)				
< 1 per week	0.251*	(0.124)	0.367**	(0.124)
1 per week	0.626***	(0.104)	0.625***	(0.123)
> 1 time per week	0.792***	(0.128)	1.008***	(0.109)
Doing useful or meaningful activities (ref. never)				
< 1 per week	0.228	(0.118)	0.461***	(0.117)
1 per week	0.500***	(0.123)	0.632***	(0.117)
> 1 time per week	0.711***	(0.118)	0.709***	(0.115)
Spending time on religion or spirituality (ref. never)				
< 1 per week	-0.114	(0.109)	-0.167	(0.120)
1 per week	-0.261*	(0.123)	-0.180	(0.120)
> 1 time per week	-0.140	(0.120)	-0.379**	(0.123)
Constant	0.502	(0.466)	0.324	(0.400)
Sigma_u	1.312		1.277	
Sigma_e	1.717		1.698	
Wald chi ²	1777.851		1390.845	
p-value	0.000		0.000	
R ²	35.2%		32.8%	
Respondents	800		800	
Vignettes	2,400		2,400	

Note: Based on a multilevel regression (GLS) with robust standard errors.

Tested with controls for design effects (i.e., vignette position and dummy variables for vignette set).

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$