On the Equivalence of Collectively and Individually Collected Responses:

Standard-gamble and Time-tradeoff Judgments of Health States

PAUL F. M. KRABBE, MSc, MARIE-LOUISE ESSINK-BOT, MD, PhD, GOUKE J. BONSEL, MD, PhD

The standard-gamble and time-tradeoff methods for valuing health states were compared in a multifactorial design with 104 student volunteers. The main aim of the experiment was to compare average individual responses with group responses for the same tradeoff tasks. Group responses were collected using an interactive voting system. The standard EuroQol system was used to describe the health states to be valued. Generalizability theory was used to analyze the results. The averages and median values of the individual responses differed from the interactively collected group values only for the more severe health states. The results showed almost identical results for the two methods, but the time tradeoff was found to be more consistent than the standard gamble. The authors conclude that 1) there is significant similarity between the results of individual and collective response modes, and 2) the standard-gamble and time-tradeoff methods produce almost equivalent values, despite their different conceptual backgrounds. In this study the aggregated individual responses and the collective responses proved to be sufficiently similar to support the validity of using aggregated individual valuations as a measure of the valuation of the group. Key words: collective judgment; standard gamble; time tradeoff; methodology; voting system; EuroQol; generalizability theory. (Med Decis Making 1996;16:120-132)

Health status has become one of the important outcome measures in the evaluation of medical interventions, in addition to the more traditionally used mortality and symptom-oriented measures. Generic (i.e., comprehensive and non-disease-specific) health-status measures commonly take the form of profiles, e.g., the Nottingham Health Profile (NHP) and the MOS Short-form 36 (SF-36). However, application of health-status data in medical evaluation research, in particular in economic analysis, requires us to go one step beyond mere profile descriptions of health status. A descriptive measure with a physical dimension (A) comprising 3 levels (1 = optimal, 2 = intermediate, 3 = worse) and a psychosocial dimension (B) with three analogous levels may, for example, generate a profile of A2B2 for patient X, and A1B3 for patient Y. However, if we are to judge whether patient X is better or worse off than patient Y, and if so, by how much, a single summary measure is required. Such summary scores may be obtained by using a valuation procedure for each health state. The resulting values may be used to combine life years with quality, both in economic evaluations (e.g., in heart transplantation, coronary artery bypass grafting, administration of ACE inhibitors) and in assessing the burdens of illness upon populations in public health modeling (e.g., the World Bank report “Investing in Health”).

In an empirical procedure designed to obtain values for health states, subjects were requested to indicate how good or how bad certain health states were for them, and to rank these health states according to their degrees of undesirability by means of methods such as the standard gamble (SG) and the time tradeoff (TTO). A major issue within the scientific field of health-status valuation is, first, whether individual value responses can be aggregated to reflect group values, and if so, how. Second, there is still considerable debate as to which method, i.e., SG or TTO, is the more valid and precise.

In this paper we compare SG and TTO data collected as group responses and individual responses.
We address the following research questions:

1. Are aggregated individual value responses comparable to the collective response of the group as a whole? We hypothesized group response to be similar to the median of the individual responses.

2. How do SG and TTO compare in terms of equivalence (validity)?

3. What are the sources of measurement error for these two elicitation methods?

To answer these questions, we conducted an experimental study with a multifactorial design (see below).

We used the EuroQol descriptive system for health status, as our study was part of the program of the EuroQol Group.

In 1987, the international and multidisciplinary EuroQol Group7,8 was established. Since that time the Group has developed a short generic instrument to describe a patient's health status based on five dimensions, each comprising three levels. Health states were generated by combining discrete levels from each dimension. Valuations for sets of EuroQol health states were elicited from population samples to enable the intended use of EuroQol in economic evaluation. The current EuroQol valuation questionnaire appeared to be suitable for postal surveys without interviewer support, and produced consistent values on a selected set of health states by means of a visual analog scale (VAS). Population surveys in the United Kingdom,7 The Netherlands,8 Sweden,10 and Norway11 showed international similarities of valuations and characteristics of the responses.15 The sensitivity of the valuations for sociodemographic variables appeared to be low,6 and little bias may be expected from the nonresponders.8 Contextual effects were minimal and reproducibility was satisfactory.13

Material and Methods

ORGANIZATION

Extensive pilot studies preceded the experiment, which included two sessions, separated by a ten-day interval. Students were recruited by handouts. The same group of 104 students participated in both sessions. For full participation they were paid approximately $65 (1993). Both sessions consisted of a sequence of predominantly valuation tasks (the results of the experiments presented here are part of a more elaborate experiment). The valuation tasks were deliberately interspaced with unrelated questionnaires, for example, on the moral acceptability of genetic manipulation, to avoid weariness and irritation due to monotony.

All participants were seated in a lecture hall with due space between them. The different methods were preceded by similar verbal explanations of the method and a few test judgments. The stimuli (the health state to be valued and, with the collective tradeoff tasks, the alternative for comparison) were always presented by slide projection. During the presentation, the instructors (GJB, MLE-B) repeated the nature of the particular tradeoff task with each stimulus, to avoid any blurring of the concepts of SG and TTO. Responses were recorded using pencil and paper for the individual tasks, and by means of an electronic interactive voting system for the collective tasks.

Conventional aggregation of the valuations of all individuals for one given health state by computation of the mean or the median is indicated by IND. This standard method was compared with a method that aimed at a genuine group or collective response by means of an interactive voting system. The collective response mode is indicated by COL.

HEALTH-STATE DESCRIPTIONS

The EuroQol concept of health status consists of five dimensions: mobility, self-care, usual activities, pain/discomfort, and mood. Each dimension has three levels, "no problem" (1), "some problems" (2), and "severe problems" (3). Health-state scenarios are produced by the selection of one level for each dimension (the best health state is thus represented by 11111). Theoretically, this set of dimensions and levels of the EuroQol instrument allows for 243 (3^5) different health-state descriptions. The EuroQol Group selected 13 of these "scenarios" as a standard set as a basis for experiments. We selected 11 states from this standard set (excluding 11111 and 33333). Previous results had indicated, however, that this set did not evenly cover the continuum between 0 and 100.5,14 Two "gaps" existed in the value range of this standard set of states, so we therefore added two other health states (12212 and 33332).

STANDARD GAMBLE

The SG concept is derived from the Von Neumann–Morgenstern (vN-M) utility gamble.15 The method16 is essentially an iterative paired comparison. The participant is presented with two alternatives, and asked to select the preferred one. One alternative offers the participant a specified certain outcome, while the other alternative offers a gamble with mutually exclusive probabilities for two reference outcomes. Conventionally, SG is operationalized as a choice between A, the certainty of being...
stationary in a specific lifelong impaired health state (the state to be valued), or B, the uncertain result of an intervention, for example a surgical procedure, with two reference outcomes. These are: a probability (p) of instantaneous and lasting improvement to perfect health, or a probability (1 − p) that the operation will fail, resulting in immediate death. By varying the p-level of the uncertainty outcome, the point of indifference between the two alternatives A and B is determined. By combining the probability values with the utility values of the reference outcomes, the utility of the stationary state is established. The method fails if the state to be valued falls out of the range covered by the two reference outcomes (for example, if in the conventional operationalization a state is valued as being worse than death). To overcome such a situation, an adaptation of the presentation of A and B is necessary.

In the present experiment, two slides were shown simultaneously to generate a collective response using SG (SG_r). The certainty of the stationary chronic health state to be valued was shown on the left, while the gamble was shown on the right. The gambling probabilities of the alternative option were varied in steps of at least 2% upwards or downwards, depending on whether the gamble or the certainty was preferred, until the participants no longer preferred one over the other. Each individual expressed his or her preference for one or another alternative by means of the voting system. The first alternative option for SG was the same for all health states to be valued, namely a 50% chance of being in the "best imaginable health state" (described as "perfect health" in other studies) and a 50% chance of being in the "worst imaginable health state." Figure 1 shows an example of the SG_r presentation. The state to be valued is shown on the left (in this example EuroQol state 21232). The alternative option on the right shows a gamble with an 88% chance of the "best imaginable health state" and a 12% (100% − 88%) chance of being in the "worst imaginable health state." It was clearly stressed to the participants that both outcomes arising from the gamble would involve chronic health states.

The description of the SG measurement process includes four specifications: the duration of the state, the exact prognosis following the state, the age of onset for the state, and whether or not the state applies to the subject himself/herself or to someone

---

*Instead of the convention of using being dead as the bad reference outcome, we used "worst imaginable health state," for reasons that are explained later. The consequences of this strategy are also discussed.
else. In addition, the subject's economic well-being should not confound the measurements. All the requirements were fulfilled in this study. Due to the fact that all the participants were students, the age of onset was similar, i.e., approximately 25 years. For the SG method all health states were chronic, so duration and prognosis were fixed and were thus the same for all participants. In The Netherlands all citizens have free access to standard medical care. Consequently there is no direct confounding with economic well-being (purchasing power for medical care is homogeneous).

For the elicitation of individual responses (SGind), the state to be valued was presented on a slide. Each individual responded conventionally by dividing a "probability pie" into two complementary parts using pencil and paper. Thus the individual presentations of these probability pies corresponded exactly to the side presentations in the collective session.

TIME TRADEOFF

The method TTO was developed by Torrance as a less complicated, conceptually different although equally sound, alternative to SG. It is based on tradeoffs similar to those of the SG, but the concept of uncertainty is omitted from the TTO method. The participant trades off survival and health status. The first alternative specifies a (suboptimal) health state with a given duration of, say, ten years. The competing alternative offers a better health status (conventionally optimal health) of shorter duration, conventionally followed by death. The point of indifference is reached by varying the duration spent in perfect health. By combining the duration value and the values of the reference outcomes, the utility of the stationary state is established.

The operationalization of TTO in the present experiment involved the following alternatives. The first option was five years in the "best imaginable health state" followed by five years in the "worst imaginable health state." Our operationalization of TTOcol is shown in figure 1. The stationary health state for the next ten years is shown on the left. The alternative in this particular example, spending 6.5 years in the "best imaginable health state" followed by the remaining 3.5 years in the "worst imaginable health state," is shown on the right. A bar, proportionally divided into two parts, indicates the numerical presentation of the numbers of years. For both options the health state would return to its present form after ten years. The years in the alternative option were varied in steps of 0.5 years.

Individual responses (TTOind), were collected from participants using pencil and paper. Each individual participant was asked to divide a "duration bar" into two parts. The collective response (TTOcol) were obtained by showing two slides simultaneously, i.e., both the stationary health state to be valued and the specified alternative. The same procedure as that used in the SG method was used to achieve the point of indifference.

Within each separate SG and TTO experiment, the states to be valued were presented in a randomized order to avoid memory effects.

THE INTERACTIVE VOTING SYSTEM

A voting system with a button-box at each seat was installed in the lecture hall. The system allowed for a dichotomous response (for example, preference for either alternative A or alternative B) and for a numerical response (1 to 10, in this study used only for identification control). The number of participants and the percentages of participants voting for the stationary state to be valued A and for the method-specific alternative option B were continuously displayed on a monitor, which was visible to the investigators only. The first slide with the alternative options from which all participants had to choose individually was the 0.5/0.5 option for SG and the 5/5 option (years) for TTO. Option B was varied until the indifference point between slide A and slide B was achieved. Indifference was defined as a situation in which the number of votes for one option was between 48% and 52%. The next option B was based on the magnitude of preference for a particular alternative, and, for all subsequent bids, on the options already offered. If, for example, during the first voting cycle, 72% of the participants voted for A and 28% for B for a particular EuroQol health-state description, option B was varied into a more desirable outcome during the next voting cycle. This involved reducing the "worst imaginable health state" percentage for the SG and reducing the number of years to be spent in the "worst imaginable health state" for the TTO. Based on the outcome of the previous assessment, a simple software program selected the next slide for alternative option B that would lead most efficiently to equivalence between options A and B. For example, the succession of presentations of option B for health state 12212 was, for the SG method: 0.5/0.5 (13% preferred B); 0.12/0.88 (58%); 0.32/0.68 (28%); 0.22/0.78 (40%); 0.16/0.84 (51%). After five bids, indifference (49% A, 51% B) was reached for this health state, and, with the last proportions, 0.16 vs 0.84, the collective valuation was taken to be 84. In the experiment, the collective response stabilized after four to six bids. We regarded the score at the indifference point as a genuine point estimator of the group response.

To investigate the test–retest reliability of the collective-response procedure, the SGcol session was repeated during the second session.
THE POSITION OF "DEAD" IN SG AND TTO

In both the SG and the TTO, the state "dead" occupies a specific position. In the SG the calibrating gamble is usually between perfect health and "dead." In Torrance's original operationalization of TTO, "dead" follows the shorter period in perfect health. In the present experiment, "dead" was replaced by the "worst imaginable health state." Similarly, for the TTO, the period in "perfect health" was followed by a complementary period in the "worst imaginable health state." These two periods together were equal to the total duration of the stationary period.

The main reason for selecting "worst imaginable health state" relates to the primary objective of our study, i.e., the comparison of the utility elicitations in a collective response mode and an individual response mode. Normally, when a health state is valued as being worse than "dead" (indicated by a preference to die immediately instead of living any number of years in the state to be valued), a modification of the SG and TTO methods is necessary. In the present study, this should be the replacement of the "worst imaginable health state" with a description of the EuroQol health state valued worse than being dead and the replacement of the stationary health state (normally the state to be valued) with "dead." However, the collective response mode precluded the use of such a complex "mirror" procedure for states worse than "dead."

A further reason for selecting "worst imaginable health state" was that in the conventional operationalization of the SG and the TTO, "dead" serves only as a benchmark. "Dead" is not an essential part of either method; neither is the use of perfect health at the other extreme. Logically, any two pairs of reference states are suitable so long as they "embrace" the state to be valued and their utility values are known. As with the SG, the TTO procedure should allow for the use of reference states other than those used conventionally. (This needs the imputation of other utility values in the final calculations; see, Llewellyn-Thomas for a test of this assumption in the SG). If for the SG and the TTO reference states other than perfect health and/or dead are used, the utilities obtained with such non-standard operationalizations are different and need rescaling factors to be comparable with utilities obtained with the conventional SG and the standard TTO (appendix A).

Furthermore, it is a matter of preference or convention to anchor the value of "dead" at 0 (zero). Inevitably this convention leads to the assigning of negative values for the worst health states, regardless of the health-description system used. In QALY calculations, negative values of health states may result in complicated computations.

Finally, we wanted to be able to compare SG and TTO data with standard EuroQol VAS data (tradeoff techniques vs a rating-scale technique). In the EuroQol standard questionnaire death is rated through a separate valuation task. With the additional measurement of the value for "dead," scores on the "healthy—worst imaginable health state" scale can be transformed to a 0–1 perfectly healthy—dead scale of values.

In order to use the results of this study, for example, for the computation of QALYs, the utility for the "worst imaginable health state" would have to be determined. This drawback does not preclude us from investigating the main aim of this study, namely the degree of comparability between the two methods and the two response modes, by statistical and psychometric methods.

STATISTICAL ANALYSIS

Means and medians were computed for the individual responses. Paired t-tests between SGind and TTOind for all health states were carried out and effect sizes were computed.

Pearson's product-moment correlation coefficient (p) and Spearman's rank-correlation coefficient (pr) are the parameters most frequently used in analyzing parallel (equivalence) data. The first is suitable for interval or ratio data, while the p coefficient is more appropriate for ordinal data or data of a higher measurement level, which do not satisfy distributional requirements for p. Despite their popularity, both are for obvious reasons essentially insufficient for testing equivalence of single responses. Nevertheless, these statistics are presented here to allow for comparison with other studies.

In order to test equivalence of single valuations, the intraclass correlation coefficient (ICC) is considered to be more appropriate. The ICC takes into account the variability due to two systematic sources of bias, i.e., a level effect between two measures and a linear transformation similar to p. Furthermore, the ICC is more flexible compared with the other two correlation coefficients as it may be estimated in designs with multiple retests or with more than two raters.

†Our use of the correlation coefficients (p, pr, ICC) for estimating coefficients of equivalence can be seen as measuring the criterion validity. In that case SG would be treated as the criterion ("gold standard") and TTO as the measuring instrument that corresponds to the criterion. By definition, the criterion must be a superior, more accurate measure of the phenomenon if it is to serve as a verifying norm. In this paper SG is seen not as a superior method for measuring preferences of health states but as a competitive method to TTO. For this reason, we have treated the two methods as equivalent, inducing us to speak of convergent validity.
Collectively and Individually Collected Responses

Conventionally, the equivalence of valuation methods has been investigated by plotting the population's mean or median health-state values for each separate method and by subsequent regression analysis (see, e.g., Torrance). Although a considerable amount of information is lost by using aggregated data for regression analysis, we present the results of this approach for comparative reasons and to study the relationship between the different methods and response modes.

An extension of the concept underlying the ICC is generalizability theory (G theory). Where ICCs deal with two-way designs (subjects × raters or occasions), which makes the ICC a special case of G theory, G theory deals with n-way designs and provides an even more flexible, practical framework for examining different sources of measurement error. G theory extends classic test theory by recognizing and estimating the magnitudes of the multiple sources (facets in G-theory language) of measurement error. In the present study, the relative contributions to "health states" by the facets "methods" and "participants" and the one- and two-order interaction terms of these two facets with "health states" are estimated within G theory and named "variance components." From these estimated variance components, a generalizability coefficient \( (\beta^2) \), analogous to the standard reliability coefficient, can be calculated for SG and TTO (appendix B).

Both ICC and G theory can be implemented within the analysis of variance (ANOVA) framework. Deyo et al. described a simplified computation method for estimating the ICC, which can be managed even with a pocket calculator. G-theory studies are not yet available as a module of conventional statistical software packages, but they require a special-purpose program or simple adaptation of existing ANOVA modules. In our study we derived the necessary estimations of variance components from 8V of the BMDP software package (details available from the first author).

Results

RESPONSE

Of the 104 participants, 46% were male. The mean age was 22 (SD = 2.48) years. All were students, 71% were medical students. SG\(_{\text{ind}}\) and TTO\(_{\text{ind}}\) each took approximately 15 minutes to complete. The complete series of bids for all 13 health states for the collective experiments, until collective indifference as defined in the section "the interactive voting system" was reached, took approximately 40 minutes each. The feasibility of these experiments was satisfactory, although at the end some participants complained of weariness. Judging from the participants' remarks and from the absence of learning effects, we regarded memory effects to be highly unlikely.

DESCRIPTIVE STATISTICS

Table 1 shows the results of the experiments after linear transformation to a uniform 0-to-100 scale

<table>
<thead>
<tr>
<th>Health State†</th>
<th>Individual Responses (Session 2)</th>
<th>Collective Response</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Median</td>
</tr>
<tr>
<td>12111 (1)</td>
<td>96.2</td>
<td>94.5</td>
</tr>
<tr>
<td>11211 (2)</td>
<td>95.5</td>
<td>92.5</td>
</tr>
<tr>
<td>11211 (3)</td>
<td>94.5</td>
<td>92.8</td>
</tr>
<tr>
<td>11211 (4)</td>
<td>93.8</td>
<td>93.5</td>
</tr>
<tr>
<td>11121 (5)</td>
<td>93.3</td>
<td>91.8</td>
</tr>
<tr>
<td>11222 (6)</td>
<td>93.3</td>
<td>86.0</td>
</tr>
<tr>
<td>12222 (7)</td>
<td>81.9</td>
<td>87.7</td>
</tr>
<tr>
<td>32211 (8)</td>
<td>79.2</td>
<td>73.1</td>
</tr>
<tr>
<td>21232 (9)</td>
<td>65.2</td>
<td>58.2</td>
</tr>
<tr>
<td>22232 (10)</td>
<td>64.6</td>
<td>60.9</td>
</tr>
<tr>
<td>33232 (11)</td>
<td>53.7</td>
<td>47.9</td>
</tr>
<tr>
<td>22233 (12)</td>
<td>51.5</td>
<td>44.9</td>
</tr>
<tr>
<td>33332 (13)</td>
<td>34.5</td>
<td>28.0</td>
</tr>
</tbody>
</table>

*All differences were significant (t-tests; p-values < 0.05) except that for health state "11112."
†The EuroQol concept of health status consists of five dimensions: mobility, self-care, usual activities, pain/discomfort, and mood. Each dimension has three levels, "no problem" (1), "some problems" (2), and "severe problems" (3). Health-state scenarios are produced by the selection of one level for each dimension (the best health state is thus represented by 11111).
‡Estimator of effect size \( d \) for continuous variables, which relates the differences in mean scores to the dispersion of the scores. An effect size \( d = 0.2 \) indicates a small effect; \( d = 0.5 \), a medium effect; and \( d = 0.8 \), a large effect.
Table 2 • Convergent Validity: the Amounts of Equivalence between Standard Gamble (Individual Response Mode) and Time Tradeoff (Individual Response Mode) as Measured by Intradass Correlation Coefficients (ICCs), Spearman Rank-correlation Coefficients (ρ), and Pearson Product-Moment Coefficients (p) for 13 Health States Based on 104 Respondents’ Valuations

<table>
<thead>
<tr>
<th>Health State*</th>
<th>ICC</th>
<th>(ρ)</th>
<th>(p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>12111 (1)</td>
<td>0.55</td>
<td>0.69</td>
<td>0.63</td>
</tr>
<tr>
<td>11211 (2)</td>
<td>0.52</td>
<td>0.64</td>
<td>0.66</td>
</tr>
<tr>
<td>21111 (3)</td>
<td>0.69</td>
<td>0.69</td>
<td>0.73</td>
</tr>
<tr>
<td>11112 (4)</td>
<td>0.55</td>
<td>0.64</td>
<td>0.60</td>
</tr>
<tr>
<td>11121 (5)</td>
<td>0.80</td>
<td>0.76</td>
<td>0.82</td>
</tr>
<tr>
<td>11122 (6)</td>
<td>0.75</td>
<td>0.74</td>
<td>0.79</td>
</tr>
<tr>
<td>12211 (7)</td>
<td>0.51</td>
<td>0.61</td>
<td>0.52</td>
</tr>
<tr>
<td>32211 (8)</td>
<td>0.63</td>
<td>0.65</td>
<td>0.66</td>
</tr>
<tr>
<td>21232 (9)</td>
<td>0.65</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td>22323 (10)</td>
<td>0.74</td>
<td>0.71</td>
<td>0.75</td>
</tr>
<tr>
<td>33232 (11)</td>
<td>0.70</td>
<td>0.72</td>
<td>0.72</td>
</tr>
<tr>
<td>22233 (12)</td>
<td>0.69</td>
<td>0.70</td>
<td>0.71</td>
</tr>
<tr>
<td>33332 (13)</td>
<td>0.65</td>
<td>0.69</td>
<td>0.68</td>
</tr>
</tbody>
</table>

Mean, all states | 0.65 | 0.67 | 0.69

*The EuroQol concept of health status consists of five dimensions: mobility, self-care, usual activities, pain/discomfort, and mood. Each dimension has three levels: "no problem" (1), "some problems" (2), and "severe problems" (3). Health-state scenarios are produced by the selection of one level for each dimension (the best health state is thus represented by 11111).

(TTO<sub>ind</sub> = 10 × score; TTO<sub>col</sub> = 10 × score; SG<sub>ind</sub> = 100 × score in degrees/360). Medians of SG<sub>ind</sub> and TTO<sub>ind</sub> are presented for appropriate comparison with the collective method. The order of presentation of the 13 health states in table 1 is arbitrarily based on the SG<sub>ind</sub> values. There is a noticeable similarity between the means and the medians of the IND values. Except for one health state, the means of the health states were statistically significantly different for SG<sub>ind</sub> and TTO<sub>ind</sub>. However, the effect size indicated that for most of the health states those differences were small.

Generally, the individually and collectively aggregated values appear to have the same structure. The similarity of SG and TTO values is observed for both response modes. Differences between individually and collectively aggregated values are more obvious. In particular, the health states 21232, 22323, 22233, and 33332 are valued differently to some extent within the two response modes. On average, the individual response to these worse states is higher than the collective response. The orders of the 13 health states for the two response modes show approximately similar results for the SG and the TTO.

MEASURES OF EQUIVALENCE

Three different correlation coefficients were computed as indices for the convergent validity between the SG and the TTO for all 13 health states, based on the individual valuations of the 104 participants. The overall means of the correlation coefficients for all health states show minor differences for the three correlation methods (table 2). The overall ICC is slightly lower than Pearson's p due to a small but distinct method effect. The test–retest ICC for the SG<sub>col</sub> was 0.97, indicating excellent reproducibility of the group responses.

Figure 2 shows the results of regression analyses between collective responses and individual responses and between the SG and the TTO. Equations for the regressions of figure 2 and for some regressions not depicted in figure 2 are presented in table 3. Regressions between collective responses and
individual responses are presented for the median values only (fig. 2A). Coefficients of determination are high, 0.97 and 0.96, respectively, for the regressions $SG_{col} - SG_{ind}$ and $TTO_{col} - TTO_{ind}$. The worse health states are valued somewhat lower in the collective sessions.

The scattergram of method-specific median values for the individual health states (fig. 2B) shows that all $SG_{ind}$ and $TTO_{ind}$ median values are on the regression line. $SG_{ind}$ is equal to $TTO_{ind}$, except for a small linear transformation (0.86), indicating that $TTO_{ind}$ valuations are a little lower than $SG_{ind}$ valuations (method effect). The coefficient of determination ($R^2$) is near 1.0. Regression between $TTO_{ind}$ and $SG_{ind}$ performed with mean values yields the same outcome. Including the regression of collective responses, $R^2$ is 0.99, despite identical values of three health states for the $SG_{col}$ method, whereas $TTO_{col}$ showed minor differences.

**SOURCES OF MEASUREMENT ERROR**

We performed G-theory analyses with the object of measurement (health states) and the two facets (participants and methods). G theory allowed us simultaneous estimations of the effects of these facets on the 13 health states, and we assumed that for each separate health state the 104 participants would produce equivalent values with both methods, $SG_{ind}$ and $TTO_{ind}$. Deviations from this assumption were seen as measurement errors. Results of this analysis, based on the seven sources of variance (H, P, M, H X P, H X M, P X M, residual), are shown in table 4. Starting with the primary effects (H, P, M), 56% of all variance was explained by the 13 health states (H). The contribution to measurement error of the participants (P) was relatively small; only 9% of variance was attributed to a systematically different valuation of the participants for all 13 health states. The systematic difference between the methods $SG_{ind}$ and $TTO_{ind}$ was negligible, i.e., $M = 1\%$. The three first-order interaction terms (H X P, H X M, P X M) were responsible for 25% of measurement error. Interaction term H X M accounted for 0% of the variance. This finding supported the preceding regression analyses, which showed that the aggregated values for $SG_{ind}$ and $TTO_{ind}$ had a perfect linear relationship with each other in this study. The largest term of measurement error was $H \times P$ (22%). This means that the major part of the measurement error resulted from some participants’ valuing specific health states differently compared with other participants and was irrespective of the method used. Nonsystematic error and the H X P X M second-order interaction (some participants valued a specific health state differently for one of the methods) were subsumed together within a small residual term (9%). A visual impression of the contributions of variance of the three sources of variance and their interactions is provided by the Venn diagram in figure 3.

Generalizability theory also allows for a closer

**Table 3**

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Prediction Based on Means</th>
<th>$R^2$</th>
<th>Prediction Based on Medians</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Individual vs collective</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard gamble</td>
<td>$SG_{col} = -50.6 + 1.55 \times SG_{ind}$</td>
<td>0.97</td>
<td>$SG_{col} = -45.8 + 1.45 \times SG_{ind}$</td>
<td>0.97</td>
</tr>
<tr>
<td>Time tradeoff</td>
<td>$TTO_{col} = -28.0 + 1.29 \times TTO_{ind}$</td>
<td>0.95</td>
<td>$TTO_{col} = -18.0 + 1.14 \times TTO_{ind}$</td>
<td>0.96</td>
</tr>
<tr>
<td><strong>Standard gamble vs time tradeoff</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Individual</td>
<td>$SG_{ind} = 10.0 + 0.91 \times TTO_{ind}$</td>
<td>1.00</td>
<td>$SG_{ind} = 15.0 + 0.86 \times TTO_{ind}$</td>
<td>1.00</td>
</tr>
<tr>
<td>Collective</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*The collective method essentially produces a median.

**Table 4**

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Sum of Squares (ss)</th>
<th>Degrees of Freedom (df)</th>
<th>Mean Square (ss/df)</th>
<th>Estimated Variance Component %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health states (H)</td>
<td>1,141,064</td>
<td>12</td>
<td>95,088.7</td>
<td>455.38</td>
</tr>
<tr>
<td>Participants (P)</td>
<td>234,150</td>
<td>103</td>
<td>2,273.3</td>
<td>73.25</td>
</tr>
<tr>
<td>Methods (M)</td>
<td>9,550</td>
<td>1</td>
<td>9,550.9</td>
<td>6.66</td>
</tr>
<tr>
<td>HP</td>
<td>455,879</td>
<td>1,236</td>
<td>368.8</td>
<td>184.42</td>
</tr>
<tr>
<td>H X P</td>
<td>3,041</td>
<td>12</td>
<td>253.5</td>
<td>1.74</td>
</tr>
<tr>
<td>H X M</td>
<td>37,869</td>
<td>103</td>
<td>367.7</td>
<td>22.73</td>
</tr>
<tr>
<td>Residual (HPM, e)</td>
<td>89,232</td>
<td>1,236</td>
<td>72.2</td>
<td>72.19</td>
</tr>
</tbody>
</table>
The first research question addressed the comparison of two response modes. The collective response was assumed to produce a value that was comparable to the median of the individual responses. The specific nature of this group response (one point estimator, no dispersion by definition) precluded statistical testing of the difference between the two response modes and also limited the use of other explanatory statistical techniques. Generally, median values of individual responses were similar to the group values, with the bad health states as exceptions. We considered the following explanations for the somewhat different response for the bad health states.

**Range compression.** Each collective valuation experiment started with a 50/50 alternative, i.e., 50% good health versus 50% bad health. Thus, the reference was different compared with the individual experiments, where a blank "probability pie" or "ten-year duration bar" was presented on paper. Under the conditions of the individual valuation experiments, the participants were not confronted with the 50/50 situation, and may consequently have been more cautious, giving low values to health states with values close to 50 and lower.

**Response shift.** In the individual response mode, bad health states were valued with considerable individual variability. The cyclic process of the collective mode may influence the consistency of individual choice. If most participants prefer the alternative state when making the first comparison, the next comparison will be between the health state to be valued and a more severe alternative state than previously. Participants who, when making the first comparison, prefer the health state to be valued, may be aware of the discrepancy between their own valuations and those of the majority of the group, and shift towards the majority judgment in the second cycle. However, the numbers of instances of this type of inconsistency for the 13 health states were computed for all individuals for each health state and method. Per health state, these values ranged from 0% to 7% for both SG<sub>ind</sub> and TTO<sub>ind</sub>. Thus, response shift can hardly have been responsible for the differences between the two response modes.

**Forced consideration.** In the individual response mode, we assumed participants would take their own health status as a latent reference due to the fact that this was a pencil-and-paper task, whereas the collective response mode more forcefully encourages participants to examine only the health state to be valued (slide presentation) and the selected alternative for several cycles. Hence, the cognitive processes underlying the two response modes are probably different, which more generally may explain the common difference between tradeoff techniques and VAS.

Addressing the second research question, we compared the SG and the TTO and examined the equivalence (convergent validity) of the individual methods using G theory as well as the more commonly used regression approach. Conventional regression analysis, based on the mean values for health states, revealed a coefficient of determination of 0.99, which is higher than the results achieved by Torrance. 18

Our experiment confirmed another phenomenon found in earlier studies, i.e., SG<sub>ind</sub> yields slightly higher valuations than does TTO<sub>ind</sub>. Risk aversion may be one of the explanations. Risk aversion is assumed to lead to a relatively higher valuation of in-

---

**Conclusions and Discussion**

Look at the reliabilities of the two valuation methods (table 5 and appendix B). Separate estimation of the generalizability coefficient yielded an internal consistency coefficient of 0.99 for the SG<sub>ind</sub> health-state valuations (based on 104 participants). This internal consistency coefficient, a specific type of reliability, stands for the precision of the valuations based on the aggregated individual responses. For TTO<sub>ind</sub> the internal consistency coefficient of the health-state valuations based on the group level was 1.0.

Additionally, one of the advantages of G theory is that it allows us to estimate internal consistency coefficients based on individual level. We estimated these coefficients, which proved, of course, to be much lower for both methods (SG<sub>ind</sub> = 0.56; TTO<sub>ind</sub> = 0.64).

**Figure 3.** Venn diagram: contributions of the different sources of variance (object of measurement H and the two sources of measurement error P and M); 56% variance uniquely attributable to the valuation of the 13 health states, 9% of variance uniquely attributable to idiosyncratic responses of the 104 participants to the health states, 1% of variance uniquely attributable to a systematic difference of valuation between standard gamble and time tradeoff; all other percentages reflect combinations of the three unique sources of variance.

Addressing the second research question, we compared the SG and the TTO and examined the equivalence (convergent validity) of the individual methods using G theory as well as the more commonly used regression approach. Conventional regression analysis, based on the mean values for health states, revealed a coefficient of determination of 0.99, which is higher than the results achieved by Torrance. 18

Our experiment confirmed another phenomenon found in earlier studies, i.e., SG<sub>ind</sub> yields slightly higher valuations than does TTO<sub>ind</sub>. Risk aversion may be one of the explanations. Risk aversion is assumed to lead to a relatively higher valuation of in-

---

**Range compression.** Each collective valuation experiment started with a 50/50 alternative, i.e., 50% good health versus 50% bad health. Thus, the reference was different compared with the individual experiments, where a blank "probability pie" or "ten-year duration bar" was presented on paper. Under the conditions of the individual valuation experiments, the participants were not confronted with the 50/50 situation, and may consequently have been more cautious, giving low values to health states with values close to 50 and lower.

**Response shift.** In the individual response mode, bad health states were valued with considerable individual variability. The cyclic process of the collective mode may influence the consistency of individual choice. If most participants prefer the alternative state when making the first comparison, the next comparison will be between the health state to be valued and a more severe alternative state than previously. Participants who, when making the first comparison, prefer the health state to be valued, may be aware of the discrepancy between their own valuations and those of the majority of the group, and shift towards the majority judgment in the second cycle. However, the numbers of instances of this type of inconsistency for the 13 health states were computed for all individuals for each health state and method. Per health state, these values ranged from 0% to 7% for both SG<sub>ind</sub> and TTO<sub>ind</sub>. Thus, response shift can hardly have been responsible for the differences between the two response modes.

**Forced consideration.** In the individual response mode, we assumed participants would take their own health status as a latent reference due to the fact that this was a pencil-and-paper task, whereas the collective response mode more forcefully encourages participants to examine only the health state to be valued (slide presentation) and the selected alternative for several cycles. Hence, the cognitive processes underlying the two response modes are probably different, which more generally may explain the common difference between tradeoff techniques and VAS.

Addressing the second research question, we compared the SG and the TTO and examined the equivalence (convergent validity) of the individual methods using G theory as well as the more commonly used regression approach. Conventional regression analysis, based on the mean values for health states, revealed a coefficient of determination of 0.99, which is higher than the results achieved by Torrance. 18

Our experiment confirmed another phenomenon found in earlier studies, i.e., SG<sub>ind</sub> yields slightly higher valuations than does TTO<sub>ind</sub>. Risk aversion may be one of the explanations. Risk aversion is assumed to lead to a relatively higher valuation of in-

---

**Conclusions and Discussion**

The first research question addressed the comparison of two response modes. The collective response was assumed to produce a value that was comparable to the median of the individual responses. The specific nature of this group response (one point estimator, no dispersion by definition) precluded statistical testing of the difference between the two response modes and also limited the use of other explanatory statistical techniques. Generally, median values of individual responses were similar to the group values, with the bad health states as exceptions. We considered the following explanations for the somewhat different response for the bad health states.
intermediate outcomes (health states); in this study, the bad health states in particular were valued higher by SG_m in comparison with TTO_m.

The individual generalizability coefficient for the TTO method of 0.64 slightly exceeded the 0.56 obtained for the SG. The inherent complexity of the SG task and the fact that people generally have difficulty with probabilistic situations may have been responsible for this result. The main method effect (SG versus TTO) was nil. This result of the G-theory analysis was equal to the results of the regression analyses, which showed predictions up to 99%. Another conclusion is that the influence of interactions between health states and methods proved to be almost zero. This means that under the conditions of this study there was no systematic difference between the orderings of the valued health states obtained with the two tradeoff methods. Moreover, it is clear that certain participants deviated from the group when valuing particular states, regardless of the method involved. The results do not support the claim for the unique conceptual position of the SG, although a specific small effect, probably due to risk, could be observed.

We conclude that under highly controlled experimental circumstances the SG and the TTO are equivalent to a large extent, despite their apparent conceptual difference. Our results can be compared with the few existing studies that have examined this issue, taking into consideration that in the latter studies the numbers of health states and/or participants have usually been small and the statistical techniques rather global. The authoritative paper of Torrance, published in 1976, reported a reliability coefficient (Pearson correlation based on replications) of 0.77 for both the SG and the TTO and a coefficient of determination (R^2) of 0.95 between the SG and the TTO. These coefficients are based on the mean values of six health states. In Torrance's study, the very bad and the very good health states were excluded, which may have improved the coefficients. Comparison of mean values obtained with the SG and the TTO for 35 disability levels by Wolfson et al. resulted in an R^2 of 0.84. Read et al. presented a Pearson correlation coefficient of 0.65 between the SG and the TTO. Their study was based on the valuation of only two health states. Hornberger et al. reported a Spearman rank correlation of 0.31 between the SG and the TTO. Their results were based on 58 individual patients' valuations of their own health.

Two issues arising from this study need some clarification. The first concerns the relatively high correlation coefficients that we found at the group level despite the considerable variation among the individual values. This can partially be explained by the fact that valuing health states is a stimulus-scaling task. There is to some extent a logical dominance of health states, which diminishes the variation among participants. For example, all participants valued 11122 as better than 33321. In contrast to attitude questionnaires, stimulus-scaling tasks frequently show high reliability coefficients. The wide range of the health states chosen as stimuli is another factor responsible for the rather high correlation coefficients.

The second issue refers to the measurement level. The purpose of most health-valuation studies is to provide a valid representation of health states along a single continuum (construct) with specified anchors for a specified population. The individual values for our type of data are neither interval data nor ordinal data but more likely to be something in between, i.e., "quasi-interval" data. Individual values of health states are at best seen as imprecise representations of the individuals' internal scales of health states. An index of this imprecision, the generalizability coefficient (internal consistency) of individual values can be seen in table 5. Information about the imprecision of individual measurements and the considerable differences among individuals is lost if we use the mean (table 5). For applications where we are interested in group values, this seems an appropriate approach. Aggregation of individual quasi-interval outcomes results in a group scale that has real interval characteristics.

We conclude that the feasibilities of the SG and the TTO were comparable in this setting, while the reliabilities or precisions of both methods based on the responses of all 104 participants were excellent. Is there any basis for preferring either method? Examination of the issue of content validity is difficult, as there is no agreed-upon "gold standard." The results of our studies have replicated the equivalence

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Estimated Variance Component</th>
<th>%</th>
<th>Estimated Variance Component</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health states (H)</td>
<td>415.41</td>
<td>56</td>
<td>497.10</td>
<td>64</td>
</tr>
<tr>
<td>Participants (P)</td>
<td>90.45</td>
<td>12</td>
<td>78.77</td>
<td>10</td>
</tr>
<tr>
<td>Residual (e)</td>
<td>238.80</td>
<td>32</td>
<td>202.23</td>
<td>26</td>
</tr>
<tr>
<td>Generalizability coefficient ((p^2))</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Per participant</td>
<td>0.56</td>
<td></td>
<td>0.64</td>
<td></td>
</tr>
<tr>
<td>For the group (n = 104)</td>
<td>0.99</td>
<td></td>
<td>1.00</td>
<td></td>
</tr>
</tbody>
</table>
of two of the best-known tradeoff methods for the valuation of health states, the SG and the TTO.

Further research should particularly address such issues as the characteristics of the two tradeoff procedures in relation to risk attitude, and other effects that influence the outcomes of these two elicitation methods as well as others. Within the field of psychometrics, a major issue is how both methods allow for the use of other non-extreme response states in the valuation task, elaborating on the axiomatic work of Llewellyn-Thomas et al.

We found considerable evidence that collective responses are comparable to the medians of individual responses for these two tradeoff techniques, although worse health states were valued lower in the collective response mode. We therefore consider it to be valid, especially for the moderate health states, to use the median of individual valuations as a measure of the valuation of the group.

The authors thank the members of the EuroQol Group, in particular Paul Kind and Stefan Björk, for their constructive comments on the design of the study and on an earlier version of the paper, and Rosalind Rabin for her refinements to the manuscript. Valuable comments were also made by two anonymous reviewers and Anne Stiggelbout, PhD.

Data for this experiment and a description of the algorithm, which uses existing data sets to select the subsequent slides in the most efficient way, are available from the authors.

References

**APPENDIX A**

**Time tradeoff (TTO)**

In the standard TTO method, subjects are asked to judge a duration \( Y \) of survival such that surviving \( Y \) years in perfect health (Q*) followed by death would be equal in preference to surviving, for example, ten years in a designated health state Q followed by death. The equation for utility in a standard TTO is:

\[
U(X, Q) = U(Y, Q^*)
\]  

where \( U(Q^*) \) ("best imaginable health state") is set at equal to 1 and \( X \) is 10 years,

\[
U(Q) = Y/10
\]  

In the present study we used "worst imaginable health state" instead; fixing \( U(\text{dead}) \) at 0 as in the standard TTO. The utility of this reference state \( U(W) \) can be set at 0 (depending on the theoretical assumption of "dead"), or it may be assumed to be known, as it is not measured directly. If assumed unknown, equation 2 becomes, for our study:

\[
U(Q) = (Y/10) + \left( \frac{10 - Y}{10} \right) \times U(W)
\]  

Equation 3 shows that the ratios \( Y/10 \) of standard TTO differ from the utilities of health states by an unknown additive factor \( U(W) \times (10 - Y)/10 \). Utilities elicited in this study are therefore not comparable with utilities elicited by standard TTO in other studies, unless \( U(W) = 0 \).

**Standard gamble (SG)**

In the conventional SG, subjects are asked to choose a probability \( P \) such that a \( P \) chance of surviving a lifetime in perfect health (Q*) and a \( 1 - P \) chance of immediate death would be equal in preference to surviving a lifetime in health state Q:

\[
U(Q) = U(P, Q^*)
\]  

when \( U(Q^*) \) ("best imaginable health state") is equal to 1,

\[
U(Q) = P
\]  

In this study "worst imaginable health state" replaced "dead" as the worst health state. An undetermined additive constant has thus been introduced, and equation 5 becomes:

\[
U(Q) = P + (1 - P) \times U(W)
\]  

If the (individual) assessment of the health state "worst imaginable health state" for the non-standard SG and the non-standard TTO are equal, equations 3 and 6 will both contain an unknown additive factor that is the same for the two methods. Comparisons and tests of equivalence with the SG and the TTO and their collective counterparts are therefore still permitted and justified.

Notice that the constant \( U(W) \) is not introducing an additional error component into the outcomes of our study, compared with the conventional SG and TTO. Setting \( U(\text{dead}) = 0 \) is an arbitrary convention; not excluding variation of individual assessments for the health state "dead" due to unique individual attitudes towards death itself.

**APPENDIX B**

**Classic test theory (CTT)**

The true-score model is the core of CTT and is expressed as:

\[
X = T + E
\]  

\( X \) is an observed score that could be envisaged as the composite of two hypothetical components: a true score (T) and an undifferentiated random-error component (E). The correlation coefficient that expresses the degree of relationship between true and observed scores is known as the reliability index:

\[
\rho_{\text{tr}} = \frac{\sigma_T}{\sigma_X}
\]  

This coefficient can be easily estimated in CTT if the data of \( k \) related items are available by computation of a coefficient of internal consistency (also referred to as homogeneity, scalability, or the like); Cronbach's \( \alpha \):

\[
\alpha = k \frac{k - 1}{(k - 1) \sum \sigma^2_i} - 1
\]  

With multi-item health status questionnaires, the goal is usually to measure a specific position of a person on a domain. The question to be answered by Cronbach's \( \alpha \) is how well the composite of all the items (\( \sigma^2_X \)) of a test is measuring the construct/domain.

Assessment of health states with an underlying natural ordering (stimulus-scaling task) yields another type of data. Instead of persons, stimuli (health states) are positioned on a scale. The question now is: how well does the
composite of the responses of all the persons to the stimuli represent the scale of these stimuli?

An adaptation of the conventional Cronbach's α is consequently required. It is not the items (health states) that are tested for their internal consistency, but the responses of the persons to the stimuli. Therefore, equation 3 shows index p (persons) instead of i (item).

Generalizability (G) theory

A much more flexible framework in comparison with CTT is G theory, fully based on analysis of variance (ANOVA). In this study we have used G theory to estimate the generalizability coefficients (internal consistencies) for SG and TTO, which resemble, in this case, equation 3 while there are two effects (health states, persons). The object of the measurement is the valuation of the health states. Moreover, we were able to estimate generalizability coefficients based on the scores of the responses of all 104 persons. The computational formulas and expected mean squares for estimating the sources of variance and generalizability coefficients for table 4 are shown in table B1.

### Table B1

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Mean Square</th>
<th>Expected Mean Square</th>
<th>Estimated Variance Component</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health states (n)</td>
<td>MS_n</td>
<td>σ²_n + n_p σ²_p</td>
<td>(\hat{\sigma}_n^2 = (MS_n - MS_p)/n_p)</td>
</tr>
<tr>
<td>Persons (p)</td>
<td>MS_p</td>
<td>σ²_p + n_h σ²_h</td>
<td>(\hat{\sigma}_p^2 = (MS_p - MS_h)/n_h)</td>
</tr>
<tr>
<td>h × p, e(c)</td>
<td>MS_e</td>
<td>σ²_e</td>
<td>(\hat{\sigma}_e^2 = MS_e)</td>
</tr>
</tbody>
</table>

\[
E_p(\text{individual}) = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_e + \sigma^2_h} \\
E_p(\text{group}) = \frac{\sigma^2_p}{\sigma^2_p/(n_p) + (\sigma^2_e/n_h) + (\sigma^2_h/n_h)}
\]

ANOVA estimates of variance components are unstable, especially with small sample sizes, and may even be negative. ANOVA is also sensitive to distributional form and unbalanced design. Our data failed only insofar as not all valuations of the health states showed normal distributions. This failure to meet the assumptions of ANOVA tends to overestimate error variances. Therefore, outcomes of the G study are conservative and may underestimate the true reliability coefficients.