Health utility bias: A meta-analytic evaluation

Jason N. Doctor[a], Han Bleichrodt[b], H. Jill.Lin[c]

[a] Department of Pharmaceutical Economics and Policy, School of Pharmacy, University of Southern California, Los Angeles, CA, U.S.A.

[b] Department of Economics and iMTA/iBMG, Erasmus University, Rotterdam, The Netherlands

[c] Department of Radiology, School of Medicine, Stanford University, Menlo Park, CA

KEYWORDS: Utility Measurement, rating scale, category scale, time tradeoff, standard gamble.

Corresponding Author: Jason N. Doctor, Ph.D., Department of Clinical Pharmacy & Pharmaceutical Economics & Policy, School of Pharmacy, University of Southern California

Acknowledgements

Jason Doctor's research was made possible by a grant from the United States Department of Health and Human Services, National Institutes of Health, National Center for Medical Rehabilitation Research (NIH-NCMRR: K01HD01221). Han Bleichrodt's research was made possible by a grant from the Netherlands Organization for Scientific Research (NWO). We would like to thank Les Lenert, Dennis Revicki, Sean Sullivan, Anne Stiggelbout, and Jose-Luis Pinto Prades in assisting with the review and identification of articles for the analysis. We also thank the two anonymous reviewers for their helpful comments on the paper.

Abstract

BACKGROUND: A common assertion is that rating scale (RS) values are lower than both standard gamble (SG) and time tradeoff (TTO) values. However, differences among these methods may be due to method specific bias. While SG and TTO suffer systematic bias, RS responses are known to depend on the range and frequency of other health states being evaluated. Over many diverse studies this effect is predicted to diminish. Thus, a systematic review and data synthesis of RS-TTO and RS-SG difference scores may better reveal persistent dissimilarities.

PURPOSE: To establish through systematic review and meta-analysis the net effect of biases that endure over many studies of utilities.

PARTICIPANTS: 2,206 RS and TTO and 1,318 RS and SG respondents in 27 studies of utilities.

DATA SOURCE: MEDLINE search from 1976 to 2004, complemented by a hand search of full length articles and conference abstracts for nine journals known to publish utility studies, as well as review of results and additional recommendations by five outside experts in the field.

DATA EXTRACTION: Two investigators abstracted the articles. We contacted the investigators of the original if required information was not available.

DATA SYNTHESIS: No significant effect for RS and TTO difference scores was observed: effect size (95% C.I.) = 0.04 (-0.02, 0.09). In contrast, RS scores were significantly lower than SG scores: Effect size (95% C.I.) = -0.23 (-.28, -0.19). Correcting SG scores for three known biases (loss aversion, framing and probability weighting) eliminated differences between RS and SG scores (effect size (95% C.I.) = 0.01 (-0.03, 0.05).

LIMITATIONS: Systematic bias in the RS method may exist but be heretofore unknown. Bias correction formulas were applied to mean not individual utilities.

CONCLUSIONS: The results of this paper do not support the common view that RS values are lower than TTO values, may suggest that TTO biases largely cancel, and support the validity of formulas for correcting standard gamble bias.

Introduction

The purpose of this paper is to establish through systematic review and meta-analysis the net effect of health utility biases that occur under different elicitation methods. Health utilities play an important role in cost-effectiveness analysis. Through health utility assessment, to each health state in the analysis a presumably unique quality weight is assigned. The standard gamble (SG), time tradeoff (TTO) and rating scale (RS) are the most common preference assessment methods for assigning such weights. However, when more than one elicitation method is employed it is often the case that more than one quality weight may be assigned to any particular health state [1, 2]. One negative implication of this is that treatment recommendations may be sensitive to the method of preference assessment [3]. Differences among health state valuation methods may be due to biases that lead to errors in measurement and result in health state utilities that are too high or too low. By seeking to understand the net effect of bias we may be in a better position to recommend certain methods that minimize the occurrence of errors.

Errors that affect measurement may be divided into two classes: 1) *systematic error* - misestimation of a measurement value that is persistent both in direction and magnitude, and, 2) *nonsystematic* error – misestimation of a measurement value that is variable in magnitude and direction. Over many observations, systematic error endures and nonsystematic error abates. We capitalize on this fact, to study within a met-analytic framework the net effect of health utility bias. As we will explain next, the TTO and the SG are affected by systematic biases and the RS by nonsystematic biases. Consequently, over many studies the bias in the RS may decrease whereas the bias in the TTO and the SG remains. By pooling the results from many studies the comparison of the TTO and the SG with the RS can, therefore, give insight in the direction of the bias in the TTO and the SG. It is important to emphasize that we do not claim that the RS is the gold standard in health utility measurement. Any single RS measurement will be affected by biases. Our point is that over many studies these biases will be reduced and this property provides a benchmark with which to compare the TTO and the SG.

Systematic Error in Health State Valuations

The TTO and SG methods are susceptible to several known effects that lead to persistent, or systematic, errors. These effects are: Loss aversion, scale compatibility, utility curvature over life duration and probability weighting. A review of these effects is beyond the scope of this paper and can be found elsewhere (see Bleichrodt [4] for review). These biases alter scores such that they deviate from a value that best characterizes preference for a health state, thus making scores too high or too low. They generally increase SG scores, have both upward and downward effects on TTO scores and are predicted to have no effect on RS scores. Table 1 provides a summary of the aforementioned known predominantly upward (+) and downward (-) causes of systematic error in SG, TTO and RS values.

INSERT TABLE 1

ABOUT HERE

Nonsystematic Error in Health State Valuation

While the RS method is not susceptible to known systematic biases, individual observations are well-known to be influenced by nonsystematic error resulting from contextual bias. With the RS method, the respondent's task is to assign categories (typically integer numbers) to health state stimuli such that succeeding categories represent equal steps in value. However, empirical research has demonstrated that characteristics of an RS response depend on the range and frequency of other health states being rated [5, 6, 7]. Figure 1 illustrates range and frequency effects for a health state with bias free health state value of 0.40.

INSERT FIGURE 1

ABOUT HERE

In each panel the x-axis represents bias free value and the y-axis denotes observed value. In the left panel, labeled “Range Effect”, one group of respondents rated the health state in context (C_1) which includes a limited range of health state values (range = 0.30 to 0.70). Because of a desire to spread responses over the full range of the response scale, the observed rating differs in C_1 than for subjects whose ratings were made in context C_2, a context with a broader range of health state values (0.0 – 1.0). In the right panel, labeled “Frequency Effect”, the health state is presented either amongst a set of health states where a preponderance have either low subjective value (C_3), or, high subjective value (C_4). By the frequency effect, observed rating response is more sensitive to changes in value when most stimuli are of similar value to the state being evaluated. An important point is that range and frequency effects produce error magnitude and direction that is specific to context; hence error is not systematic but changes with context. Schwartz [8] applied range-frequency theory to explain with great precision contextual bias in RS scores reported elsewhere [5]. Robinson et al. [6] confirmed this finding in a separate experiment. Pollack [9, 10] demonstrated convincingly that rating scales could be unbiased when contextual factors were varied iteratively over many experiments i.e., Pollack [9, 10] identified and subsequently manipulated bias effects to neutralize bias. The nonsystematic nature of rating scale context bias suggests that over many naturally occurring studies rating scale bias may decrease in size.

Whether or not SG or TTO values are influenced by nonsystematic factors like context has received much less attention. Robinson et al. [6] found in a context manipulation experiment that SG values were much less susceptible to context effects than were RS values. We are unaware of any studies examining context effects and TTO responses.

Comparing RS, TTO and SG Values

Empirically, RS, TTO and SG values do not appear to agree. A common assertion is that RS values are lower than TTO and SG values [1, 2]. However, given that the RS is subject to a context bias, one may not conclude from any single study, that RS values are lower or higher than TTO or SG values. This caveat applies even when no explicit context is given, in particular, when respondents rate only their current health. Birnbaum [11] has shown that when not given an explicit context, respondents choose their own contexts and choose different ones for different stimuli. He was in fact able to show through a between-subjects experiment that the number "9" achieved a higher largeness rating than the number "221". Presumably, "9" is large in the context of one digit numbers and "221" is small in the context of three digit numbers. Such an effect appears not easily alleviated by explicit use of anchors at points along the rating scale [11,12]. Hence, conclusions about relative value differences between TTO (or SG) and RS drawn from data collected within any single study where not every respondent rated the same health states are also not likely trustworthy. Only by comparing RS values against TTO (or SG) values in explicit contexts, across many studies and administered within-subject is it likely that context effects will diminish. In this paper, using a meta-analytic approach, we address the question of the overall effect of bias on TTO and SG scores. We capitalize on the fact that while the TTO and SG are susceptible to biases that result in systematic error in health state value, another method, the rating scale (RS) is susceptible to contextual effects that are nonsystematic across studies. Hence, while nonsystematic error diminishes when rating scale data are aggregated over many studies, systematic TTO and SG method error should persist.

Methods

Search Strategy and Inclusion Criteria

We searched (with no language restrictions) for all reports where RS and the TTO measures, or, SG and TTO measures were given to the same subjects evaluating the same health state at any one measurement interval. We performed a MEDLINE search using the following queries in all fields: 1) `(rating scale OR category scale OR visual analogue scale OR visual analog scale) AND (time tradeoff OR time trade-off)`, and 2) `(category scale OR rating scale OR visual analogue scale OR visual analog scale) AND standard gamble`. These searches were thought to be general enough to contain, as a smaller subset, as many studies as possible within our inclusion criteria (listed below). The search period was January 1st of 1976 through December 31st of 2004. We also completed a second manual search of 9 journals that are well-known to publish health utility data (see Table 2).

INSERT TABLE 2

ABOUT HERE

This second search was conducted to: 1) identify articles possibly missed by the MEDLINE search and, 2) extract results from abstracts published from conference proceedings printed in a subset of the journals listed in Table 2. The latter was done to avoid publication bias. When findings reported in an abstract were later published as a full-length article, only the data from the full length article were used in the meta-analysis. We complemented our search by reviewing the reference lists from original research and review articles. Finally, we circulated the list of studies we found to five experts in the field to see whether they could come up with more studies. Experts were included if they had been a lead or senior author on a paper found on the list generated by our search methods. Four experts accepted and one declined on the grounds that she had not worked in the area for some time. The expert who declined did recommend a well-known replacement who agreed to serve as the fifth expert.

Inclusion criteria were: 1) studies that elicited, for the same set of subjects, multiple methods of utility assessment, 2) multiple methods had to include the RS method along with either the SG or TTO methods, 3) all subjects had to receive the same health state descriptions, 4) reported utility scores had to be elicited, and could not be predicted from formulas or multi-attribute questionnaires (e.g., EQ-5D, Health Utilities Index, or Quality of Well-Being Scale), and 5) for TTO studies duration in current health had to exceed 5 years due to a documented unwillingness to trade time over short durations [13]. After consultation with experts a fifth inclusion criteria was added: Health states had to be evaluated by respondents as "better than death". Studies that did not meet the inclusion criteria were excluded. We note that by our third criterion, health state descriptions had to be hypothetical and could not reflect an individual's unique current health description; nor could the health state choice set be manipulated in a between-subjects experiment.

We contacted the investigators of the original studies if information was required to establish inclusion criteria or information on utility for health state was not available in the published reports. Missing data that could not be resolved by attempts to contact the authors were median imputed. Two investigators abstracted the articles. They resolved disagreements by consensus.

Statistical Analysis

Using the rmeta package within the statistical computing language R [14], we conducted two meta-analyses on effect size data over the aforementioned studies. The primary meta-analysis compared within-subject effect sizes for RS and TTO score differences. A secondary meta-analysis compared within-subject effect sizes for RS and SG score differences. A standard effect-size (d) estimate for within-subject score differences was used [15]:

$$d = \frac{M_{RS} - M_z}{S.D._{diff}}, \quad [1]$$

where M_{RS} is the mean RS score, M_z is the mean score for the competing method (either SG or TTO) and $S.D._{diff}$ is the standard deviation of the difference scores between the RS and competing method. In our case, the effect size estimates the average score difference (between two utility elicitation methods) relative to the variability in task performance in the population. In order to compute standard deviation of difference scores, an estimate of the population correlation between RS and TTO and RS and SG ratings is needed [16]. While several correlation statistics on these rating methods have been given in the early QALY literature (see [17-19]), Nickerson [20] has differentiated among several types of correlations between utility elicitation methods and recommends use of a mean within-respondent correlation in any analysis postulating that psychological processes affect response (p.494). Such is the case with our current analysis which considers that responses are affected by psychological biases. Two papers provide appropriate (mean within-respondent) correlations for our meta-analytic purposes they are Kartman et al. [21] and Krabbe et al. [22]. With respect to the mean within-respondent correlation, r, between RS and TTO scores, Krabbe et al. [22] report this value as r = 0.23, whereas Kartman et al. [21] report a value of r = 0.25. For this analysis, we report our results under the assumption of the middle value between these two, r = 0.24. For the RS and SG difference score meta-analysis, we report our results under the assumption that r = 0.19. This is half-way between the value reported by Krabbe et al. [22] r = 0.22, and that of Kartman et al. [21], r = 0.16. For each analysis we also ran meta-analyses under the range of standard error assumptions as given by the range of published correlations between measures. This was done to determine the robustness of our findings. Context bias associated with the rating scale depends on the specific study methods, but is statistically independent across studies. Therefore, to preserve this independence assumption an average effect size computed over utilities elicited for multiple health states *within* study served as the dependent variable.

We chose to conduct random-effects (as opposed to fixed-effects) analyses of data because rating scale context bias would naturally produce statistically heterogeneous effect sizes across studies. The random-effects model incorporates a between study component of variance to address heterogeneity, whereas a fixed-effects model does not. An effect size and confidence interval plot as well is given for the primary analysis.

In addition to analysis on raw standard gambles, we conducted two meta-analyses on corrected scores. A correction formula that adjusts for the effects of bias associated with prospect theory [23] (loss aversion, framing and probability weighting) has been proposed [24] and applied elsewhere [25]. The first formula we used corrected for only probability weighting [26, 27]. We applied a one-parameter weighting function as given in Tversky & Kahneman [23] to standard gamble scores (with the standard assumption that $\gamma = .61$ (see p. 309, Equation 6 [23]). This gives a standard gamble utility corrected for probability weighting. The second analysis utilized the following table [24]:

INSERT TABLE 3

ABOUT HERE

In addition to correcting for probability weighting, this table of values corrects for loss aversion and framing effects. This table has been used successfully to correct SG bias in other work [24].

Finally, an evaluation of study quality was considered. We evaluated the extent to which studies we examined adhered to reporting standards for studies of utilities. Each study received a quality score based on adherence to ten components of reporting standards given in Table 1 of

Stalmeier et al. [28]. Quality score was computed as the weighted sum of these ten components and scaled so that a score of 100 reflected complete adherence and a score of 0 reflected complete non adherence. Component weightings were determined by mean expert importance ratings reported in Stalmeier [28, Table 1 p.206]. We evaluated the correlation of study quality with effect size, standard error and year of publication. We also employed quality scores as weights to determine if this influenced meta-analytic findings.

Results

With regard to the RS and TTO meta-analysis, we identified 4 articles from systematic reviews, the MEDLINE search yielded 139 results, of these 13 met the inclusion criteria and were not already identified in the systematic review articles. An additional 2 studies (conference presentations) were included from a hand search of the journals in Table 1 and known review articles. Experts were not able to identify any additional RS and TTO studies that met our criteria. A total of 19 studies were used for the RS and TTO meta-analysis. With respect to the RS and SG meta-analysis, we identified 7 articles from systematic reviews, the MEDLINE search yielded 150 results, of these 5 met the inclusion criteria and were not already identified in the systematic review articles. An additional 3 studies (conference presentations) were included from a hand search of the journals in Table 2. After circulating our list to experts, they were able to identify one additional study that met our inclusion criteria and which was added. A total of 16 studies were used for RS – SG meta-analysis. We note that, as would be expected, studies utilized in the RS-TTO and RS-SG meta-analyses were not mutually exclusive. A total of 27 studies were used as data. Of these studies, eleven collected only RS and TTO responses [29-39], nine collected only RS and SG responses [40-48] and seven collected both RS, TTO and SG responses [17, 19, 49-53].

Results indicate no significant effect for RS and TTO difference scores: effect size (95% C.I.) = 0.04 (-0.02, 0.09). Figure 2 shows the plot of confidence intervals centered on effect size (x-axis) for each study. The "X" indicates an overall effect, the line through it is the confidence interval. While there is a small overall effect of 0.04, the confidence interval around this estimate crosses 0.0. These results were robust over the range of reported correlations between RS and TTO values.

INSERT FIGURE 2

ABOUT HERE

As mentioned previously, a quality score was determined by the extent to which studies adhered to published reporting criteria for studies of utility [28]. Adherence was weighted by published expert ratings of importance [28] and normalized so that a score of 100 indicates total adherence in reporting and a score of zero indicates total non adherence. Quality scores for RS-TTO studies ranged between 21.0 and 95.7. The mean (± S.D.) importance weighted quality score for RS-TTO studies was 64.7 (± 17.9). An evaluation of Pearson's product-moment correlations indicated that quality score was not significantly correlated with effect size (r = 0.23, p = n.s.), standard error (r = -.28, p = n.s.) or year of publication (r = 0.0, p = n.s.). Adding quality weights did not significantly influence meta-analytic results in that the confidence interval for RS–TTO effect size still crossed zero.

In contrast, the meta-analysis on RS and SG values indicated that RS scores were significantly lower than SG scores: effect size (95% C.I.) = -0.23 (-.28, -0.19). These results were robust to over the range of reported correlations between RS and SG values. Figure 3 shows the plot of confidence intervals centered on effect size estimates (x-axis) for each of the 16 studies included in the analysis.

INSERT FIGURE 3

ABOUT HERE

Again, The "X" indicates an overall effect, the line through it is the confidence interval. The effect is sizeable and the confidence interval around the estimate does not cross zero.

Quality scores for RS-SG studies also ranged between 21.0 and 95.7. The mean (± S.D.) importance weighted quality score for RS-TTO studies was 59.4 (± 19.3). An evaluation of Pearson's product-moment correlations indicated that quality score was not significantly correlated with effect size (r = 0.22, p = n.s.), standard error (r = -.20, p = n.s.) or year of publication (r = -0.20, p = n.s.). Adding quality weights did not significantly influence meta-analytic results in that the confidence interval for RS-SG effect size did not overlap with 0.0 and registered SG scores as consistently higher than RS scores.

The meta-analyses on corrected standard gamble scores revealed that the probability weighting correction was effective in reducing SG and RS difference, but left a very small measurable difference between SG and RS scores (effect size (95% C.I.) = -0.09 (-0.13, -0.05)). The correction adjusting for loss aversion, framing and probability weighting (see Table 1, p. 1505 in Bleichrodt et al. [24]) eliminated differences altogether, (effect size (95% C.I.) = 0.01 (-0.03, 0.05).

Discussion

An early influential review of the health utility field suggested that TTO scores were higher than RS scores [1]. This assertion was based on the best available data at the time and has remained largely unchallenged. However, 15 years later we find that contrary to this notion that RS scores are lower than TTO scores, RS and TTO scores are about equal when data are examined systematically over many within-subject studies. This may indicate that when RS context bias diminishes, value measurement becomes consistent and TTO and RS values agree. Another interpretation of this result is that, competing systematic TTO biases may cancel out. Hence, TTO scores may be relatively unbiased within a study. In either case, the discrepancy between our result that TTO and RS agree and the previous result that TTO scores exceed RS scores is likely due to diminishing RS context bias unique to the meta-analytic approach we used. In contrast, and as expected, SG biases, which are generally upward, result in higher scores than when the same individuals rate the same health states using the RS method. The disparity between SG and RS disappears when SG scores are corrected for probability weighting, framing and loss aversion.

There are a few caveats to our results that deserve discussion. First, it is important to realize that our results do not suggest that RS and TTO scores are comparable or interchangeable within a study. Hence, our study should not be interpreted as offering support for the use of the RS in economic evaluations of health care. RS scores vary substantially within a study due to context effects unique to the study. Our findings show that when evaluated systematically across many studies, TTO scores do not appear to be higher than RS scores. We are inclined to interpret this as evidence that the systematic biases in the TTO tend to cancel. Second, while no systematic RS biases are known, it is possible that one or more do exist [54], which could threaten the interpretation that TTO scores overall do not exhibit a directional bias. However, given our current state of knowledge we can be confident that TTO directional bias is not large in comparison to the directional bias exhibited by the SG method. Third, with respect to our analysis of standard gamble corrections, the fundamental data element in our study is mean score for health state; it is not guaranteed that a transformed mean score will equal a mean of transformed scores. However, transformed mean scores will approach mean transformed scores as standard errors approach zero. In most cases, standard errors were low in the studies we evaluated. Fourth, other features of elicitation methodologies such as reliability, validity and responsiveness to change are important but beyond the scope of this paper.

A large body of literature assumes that because the SG is rooted in the axioms of expected utility theory and is the only scaling method that includes an element of risk inherent in most medical decisions, the SG represents the reference standard and that other methods (e.g., the RS) should be adjusted to match SG scores [54]. We do not agree with this point of view. There is much evidence to suggest that expected utility is not the correct descriptive model (i.e., it may not characterize observed preference behavior very well). When decision makers deviate from expected utility, the SG method will generally yield biased utilities. For this reason, our method of adjusting scores does not entrust the SG method with preeminence over other methods and does not relate RS or TTO scores via mapping them to SG as is commonly done.

A basic assumption of this paper is that different methods should produce the same utilities. A practical rationale for this assumption is that if differences occur then the outcome of an economic evaluation will depend on the method used. In the absence of a gold standard for health utility measurement this is undesirable. Such an assumption is not universally held. One theory that became popular in the 1970s and 1980s, contends that risky utility (e.g., SG) and riskless value (e.g., TTO and RS) may differ by an increasing nonlinear transformation when risk aversion is considered [55]. In present day, this theory has become less popular for two reasons. First, it does not permit violations of expected utility theory, which are widely observed [56]. Second, it leads to serious problems in reconciling attitudes toward risk of small and large stakes losses [57]. For these reasons risk behavior is now primarily modeled, at its source, as attitude toward chance (via nonlinear transformation of probabilities) and through the acknowledgement that decision makers are averse to losses [23]. For an excellent discussion of how this modern approach moves toward a unified notion of utility, one that has meaning prior to risk and not visa versa, see Wakker [58]. Empirical studies have shown that when attitude toward chance and loss aversion are considered, differences between riskless and risky utility tend not to prevail [59, 60, 61].

The findings of this study have implications for cost-effectiveness analysis. In cost-effectiveness analysis, health utility assessment is carried out so that quality weights can be assigned to health states in the analysis. As demonstrated here and elsewhere, methods and procedures applied to the same health state often result in values that are inconsistent with respect to each other. Inconsistencies mean that more than one quality weight can be assigned to any particular health state. However, the valid application of CEA requires that one and only one quality weight be assigned to any particular health state. The present study is part of a growing number of studies suggesting that biases that lead to differences between measures can be reduced or eliminated. Biases appear to distort preferences in lawful and thus correctable ways, with corrections yielding greater consistency across methods. The findings of this paper suggest that standard gambles may need to be corrected for probability weighting bias. Loss aversion and framing effects may also be of concern with the standard gamble. In contrast, the findings of this paper do not support a net directional systematic TTO bias and give further support to the use of raw TTO values in cost-effectiveness analysis. Finally, while RS contextual bias may diminish over many studies, unless contextual bias is manipulated and neutralized within an experiment it is likely to adversely influence ratings in individual studies.

References

1. Froberg DG, Kane RL. Methodology for measuring health-state preferences. II. Scaling methods. J Clin Epidemiol. 1989; 42: 459-471.

2. Drummond MF, O'Brien B, Stoddart G, Torrance GW. *Methods for the Economic Evaluation of Health Care Programmes.* 2nd ed. Oxford, UK: Oxford University Press; 1997.

3. Elkin, E.B., Cowen, M.E., Cahill, D., Steffel, M. & Kattan, M.W. Preference Assessment Method Affects Decision-Analytic Recommendations: A Prostate Cancer Treatment Example. Med Decis Making 2004; 24: 504-510.

4. Bleichrodt H. A New Explanation for the Difference Between SG and TTO Utilities. Health Economics. 2002; 11: 447-456.

5. Bleichrodt H, Johannesson M.. An Experimental Test of a Theoretical Foundation for Rating Scale Valuations. Med Decis Making. 1997; 17: 208-216.

6. Robinson A., Loomes, G. Jones-Lee, M. Visual analog scales, standard gambles, and relative risk aversion. Med Decis Making. 2001; 21: 17-27.

7. Parducci A. Category judgment: A Range-frequency model. Psychol Rev. 1965; 75: 407-418.

8. Schwartz A. Rating scales in context. Med Decis Making. 1998; 18: 236.

9. Pollack I. Iterative techniques for unbiased rating scales. Q J Exp Psychol. 1965; 17: 139-148.

10. Pollack I. Neutralization of stimulus bias in the rating of grays. J Exp Psychol. 1965; 69: 564-578.

11. Birnbaum, M.H. How to Show That 9 > 221: Collect Judgments in a Between-Subjects Design, Psychological Methods. 1999; 4, 243-249.

12. Birnbaum, M.H. Using contextual effects to derive psychophysical scales. *Perception & Psychophysics*; 1974, 15, 89-96.

13. McNeil B.J., Weichselbaum R, Pauker, S. Speech and survival: Tradeoffs between quality and quantity of life in laryngeal cancer. New England Journal of Medicine. 1981; 305, 982-987.

14. Ihaka R, Gentlemen R. R – A language for data analysis and graphics. J Comp Graph Stat. 1996; 5: 299-314.

15. Morris SB, DeShon RP. Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. Psychol Methods. 2002; 7: 105–125.

16. Howell DC. Fundamental Statistics for the Behavioral Sciences, Second Edition. Boston: PWS-Kent, 1989.

17. Torrance GW. Social preferences for health states: an empirical evaluation of three measurement techniques. Socio-Econ Plan Sci. 1976; 10: 129-136.

18. Wolfson AD, Sinclair AJ, Bombardier C, McGeer A. Preference measurements for functional status in stroke patients: interrater and intertechnique comparisons. In: Kane RL, Kane RA (eds). Values and Long Term Care. Lexington, MA: Lexington Books, 191-214, 1982.

19. Read JL, Quinn RJ, Berwick DM, Fineberg HV, Weinstein MC Preferences for health outcomes: Comparison of assessment methods Med Decis Making. 1984; 4:315-329.

20. Nickerson CE. Assessing convergent validity of health-state utilities obtained using different scaling methods. Med Decis Making. 1999; 19: 487-496.

21. Kartman B, Gatz G, Johannesson M. Health state utilities in gastroesophageal reflux disease patients with heartburn: a study in Germany and Sweden. Med Decis Making. 2004; 24: 40-52.

22. Krabbe PFM, Essink-Bot M, Bonsel, GJ. The comparability and reliability of five health-state valuation methods, Soc Sci Med. 1997; 45: 1641-1652.

23. Tversky A, Kahneman D. Advances in prospect theory: Cumulative representation of uncertainty. J Risk Uncertain. 1992; 5: 297-323.

24. Bleichrodt, H, Pinto JL & Wakker. PP. Making Descriptive Use of Prospect Theory to Improve the Prescriptive Use of Expected Utility. Manage Sci. 2001; 47: 1498-1514.

25. van Osch SMC, Wakker PP, van den Hout WB, Stiggelbout AM. Correcting biases in standard gamble and time trade-off utilities. Med Decis Making. 2004; 24: 511-517.

26. Prelec D. The probability weighting function. Econometrica. 1998; 66: 497-527.

27. Wakker P, Stiggelbout A. Explaining distortions in utility elicitation through the rank-dependent model for risky choices. Med Decis Making. 1995; 15: 180-186.

28. Stalmeier, P.F. Goldstein, M.K., Holmes, A.M., Lenert, L., Miyamoto, J., Stiggelbout, A.M., Torrance, G.W., Tsevat, J. What should be reported in a methods section on utility assessment? Medical Decision Making, 2001; 21, 200-207.

29. Tsevat J, Solzan JG, Kuntz KM, Ragland J, Currier JS, Sell RL, et al. Health values of patients infected with human immunodeficiency virus: Relationship to mental health and physical functioning. Med Care. 1996; 34: 44-57.

30. Elkin EB, Cowen ME, Cahill D, Steffel M, Kattan MW. Preference assessment method affects decision-analytic recommendations: A prostate cancer treatment example. Med Decis Making. 2004; 24: 504-510.

31. Merlino LA, Bagchi I, Taylor TN, et al. Preference for fractures and other glucocorticoid-associated adverse effects among rheumatoid arthritis patients. Med Decis Making. 2001; 21: 122-132.

32. Mackeigan LD, O'Brien BJ, Oh PI Holistic versus composite preferences for lifetime treatment sequences for type 2 diabetes. Med Decis Making. 1999; 19: 113-121.

33. Sculpher M, Michaels J, McKenna M, Minor J. A cost-utility analysis of laser-assisted angioplasty for peripheral arterial occlusions Int J of Technol Assess Health Care. 1996; 12: 104-125.

34. Sanderson K, Andrews G, Corry J, Lapsley H. Using the effect size to model change in preference values from descriptive health states. Qual Life Res. 2004; 13: 1255-1264.

35. Bosch JL, Hunink MGM. Comparison of the health utilities index mark 3 (HUI3) and the EuroQol EQ-5D in patients treated for intermittent claudication. Qual Life Res. 2000; 9: 591-601.

36. Ackerman SJ, Beusterien KM, Mafilios MS, Wood MR. Measuring preferences for living in U.S. States: A comparison of the rating scale, time trade-off, and standard gamble Acad Radiol. 1998; 5(Suppl 2): S291-S296.

37. Zug KA, Littenberg B, Baughman RD, et al. Assessing the preferences of patients with psoriasis - a quantitative, utility approach. Arch Dermatol. 1995;131: 561-568.

38. Daly E, Gray A, Barlow D, Mc Pherson K, Roche M, Vessey M. Measuring the impact of menopausal symptoms on quality of life. BMJ. 1993; 307: 836-840.

39. Schwarzinger M, Stouthard MEA, Burstrom K, Nord E, European Disability Weights Group. Cross-national agreement on disability weights: the European Disability Weights Project. Popul Health Metr. 2003; 1: 9-19.

40. Revicki DA, Wu AW, Murray MI, Change in clinical status, health status, and health utility outcomes in HIV-infected patients. Med Care. 1995; 33: AS173-AS182.

41. Llewellyn-Thomas JA, Thiel EC, McGreal MJ. Cancer patients' evaluations of their current health states: The influence of expectations, comparisons, actual health status, and mood. Med Decis Making. 1992; 12:115-122.

42. Revicki DA. Relationship between health utility and psychometric health status measure. Med Care. 1992; 30: MS274-MS282.

43. Llewellyn-Thomas H, Sutherland HJ, Tibshirani R, Ciampi A, Till JE, Boyd NF. Describing health states. Methodologic issues in obtaining values for health states. Med Care 1984; 22: 543-52.

44. Sullivan SD, Lew DP, Devine EB, et al. Health state preference assessment in diabetic peripheral neuropathy. Pharmacoeconomics. 2002; 20: 1079-1089.

45. Revicki DA, Shakespeare A, Kind P. Preferences for schizophrenia-related health states: a comparison of patients, caregivers and psychiatrists. Int Clin Psychopharmacol. 1996;11: 101-108.

46. Lenert LA, Ziegler J, Lee T, Sommi R, Mahmoud R. Differences in health values among patients, family members, and providers for outcomes in Schizophrenia. Med Care. 2000; 38: 1011-1021.

47. Prades JP. Is the person trade-off a valid method for allocating health care resources? Health Econ. 1997; 6: 71-81.

48. Lyerly AD, Myers ER, Faden RR, The ethics of aggregation and hormone replacement therapy. Health Care Anal. 2001; 9: 187-221.

49. Bosch JL, Tetteroo E, Mali WP, Hunink MG. Iliac arterial occlusive disease: Cost-effectiveness analysis of stent placement versus percutaneous transluminal angioplasty: Dutch Iliac stent Trial Study Group. Radiol. 1998; 208:641-648.

50. Richardson J. Economic Assessment of Health Care: Theory and Practice. Australian Econ Rev. 1991; 93: 4-21.

51. Neumann PJ, Blumenschein K, Zillich A, Johannesson M, Kuntz KM, Chapman RH, Weiss St, Kitch BT, Fuhlbrigge AL, Paltiel AD Relationship Between FEV1% Predicted and Utilities in Adult Asthma. Society for Medical Decision Making, 22nd Annual Meeting, Cincinnati, Ohio, 2000.

52. Richardson J, Hall J, Salkeld G The measurement of utility in multiphase health states. Int J of Technol Assess Health Care. 1996; 12: 151-162.

53. Jonsson B, Horisberger B, Bruguera M, Matter L. Cost-benefit analysis of hepatitis-B vaccination Int J Technol Assess Health Care. 1991; 7: 379-402.

54. Torrance GW, Feeny D, Furlong, W. Visual Analog Scales: Do They have a Role in the Measurement of Preferences for Health States? Medical Decision Making. 2001; 21: 329-334.

55. Dyer, J. S., & Sarin, R. K. (1982). Relative risk aversion. Management Science, 28, 875-886.

56. Starmer, C. (2000). Developments in Non-expected Utility Theory: The Hunt for a Descriptive Theory of Choice under Risk. Journal of Economic Literature, 38, pp. 332-382.

57. Rabin, M. Risk Aversion and Expected-Utility Theory: A Calibration Theorem. Econometrica,. 2000; 68, 5, 1281-1292.

58. Wakker, P. 1994, Separating marginal utility and probabilistic risk aversion. Theory and Decision. 1994; 36, 1-44.

59. Stalmeier, P.F.M. & Bezembinder, T.G.G. The Discrepancy between Risky and Riskless Utilities: A Matter of Framing? Medical Decision Making. 1999; 19, 435-447.

60. Abdellaoui, M. Barrios, C. & Wakker, P.P. Reconciling Introspective Utility with Revealed Preference: Experimental Arguments Based on Prospect Theory. Journal of Econometrics. 2007; 138, 356-378.

61. Attema, A.E., Bleichrodt, H. & Wakker, P.P. Measuring the utility of life duration in a risk-free versus risky situation. 2006. Working paper, Erasmus University.

Table 1. Known predominantly upward (+) and downward (-) causes of systematic error in SG, TTO and RS values

Type of Effect	**SG**	**TTO**	**RS**
Loss Aversion	+	+	No Effect
Scale Compatibility	Ambiguous	+	No Effect
Utility Curvature	No Effect	-	No Effect
Probability weighting	+	No Effect	No Effect

Table 2. Journals searched by hand for full-length articles and or conference abstracts possibly missed by MEDLINE search

Journal Title	**Search Interval**
Health Economics	1984 - 2002
Health Policy	1984 – 1989
Health Policy in Amersterdam and Netherland	1989 – 2000
International Journal of Technology Assessment in Health Care	1985 – Present
Journal of Health Economics:	1984 – 2002
Medical Care	1978 – Present
Medical Decision Making	1981 – Present
Quality of Life Research	1993 – Present
Pharmacoeconomics	1992 – Present

Table 3. Corrected standard gamble utilities as proposed by Bleichrodt et al. [24] for standard gamble elicitations between 0.00 and 0.99. Row headings represent tenths, column headings hundredths of the uncorrected standard gamble score and table entries are corrected scores, e.g., the corrected utility for a standard gamble of 0.15 is 0.123 (underlined).

	0.00	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	0.000	0.025	0.038	0.048	0.057	0.064	0.072	0.078	0.085	0.091
0.1	0.097	0.102	0.108	0.113	0.118	0.123	0.128	0.133	0.138	0.143
0.2	0.148	0.152	0.157	0.162	0.166	0.171	0.176	0.180	0.185	0.189
0.3	0.194	0.199	0.203	0.208	0.213	0.217	0.222	0.227	0.231	0.236
0.4	0.241	0.246	0.251	0.256	0.261	0.266	0.271	0.276	0.281	0.286
0.5	0.292	0.297	0.303	0.308	0.314	0.320	0.325	0.331	0.337	0.343
0.6	0.350	0.356	0.363	0.369	0.376	0.383	0.390	0.397	0.405	0.412
0.7	0.420	0.428	0.436	0.445	0.454	0.463	0.472	0.481	0.491	0.502
0.8	0.512	0.523	0.535	0.547	0.560	0.573	0.587	0.601	0.617	0.633
0.9	0.650	0.669	0.689	0.710	0.734	0.760	0.789	0.822	0.861	0.911

Figure 1. Observed rating responses for a hypothetical health state with "context free" value of 0.40 presented in four between-subject contexts: Restricted stimulus range (C_1), broad stimulus range (C_2), positively skewed stimulus set (C_3), and negatively skewed stimulus set (C_4). The left panel shows a range effect on observed rating response (C_1 versus C_2), the right panel shows a frequency effect on observed rating response (C_3 versus C_4).

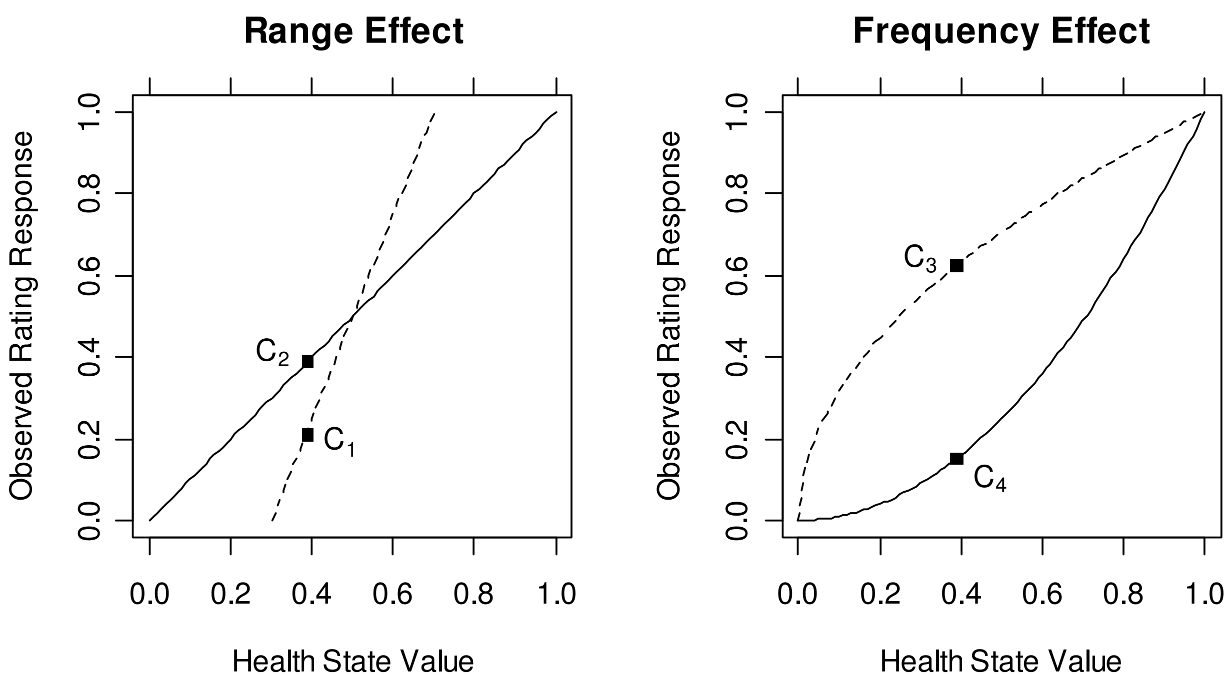

Figure 2. Plot of RS and TTO difference score effect sizes and confidence intervals for 19 studies.

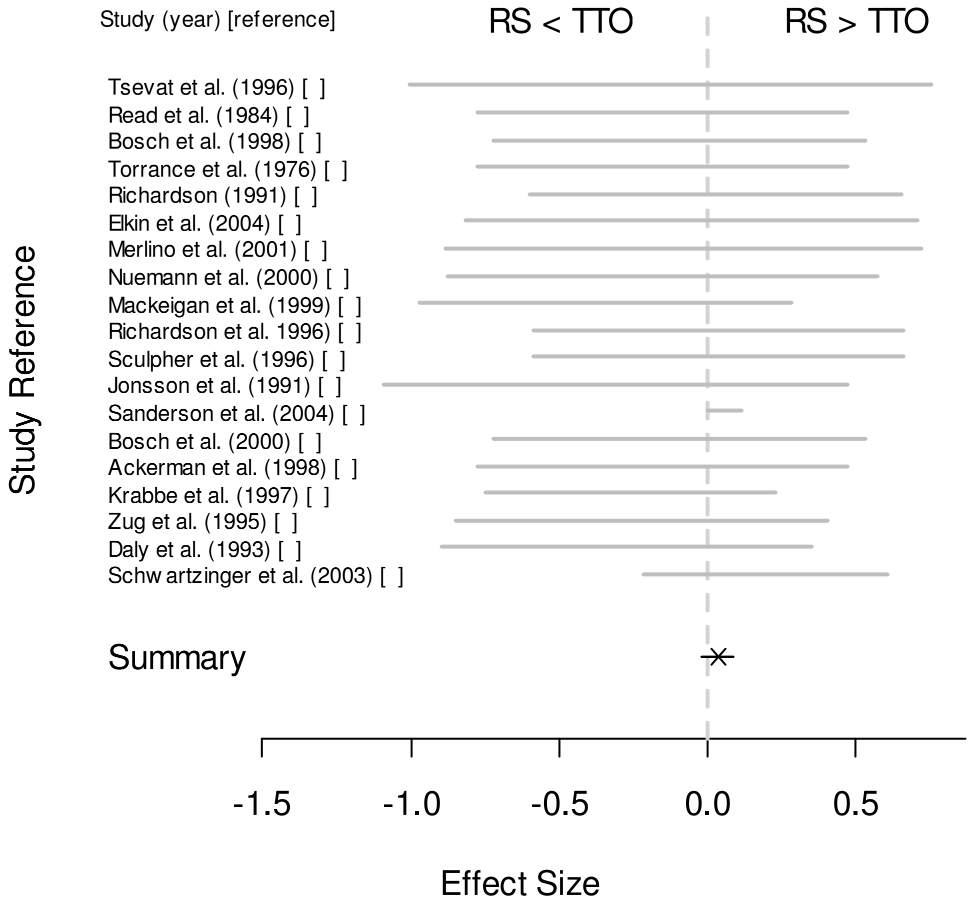

Figure 2. Plot of RS and SG difference score effect sizes and confidence intervals for 16 studies.