International Journal of Technology Assessment in Health Care, **19:1** (2003), 135–147.

PAPER STANDARD GAMBLE

A Paper-based Measure of Standard Gamble Utility for Current Health

Phillip L. Ross
Memorial Sloan-Kettering Cancer Center

Benjamin Littenberg
University of Vermont

Paul Fearn
Peter T. Scardino
Pierre I. Karakiewicz
Michael W. Kattan
Memorial Sloan-Kettering Cancer Center

Abstract

Objectives: To develop and validate a paper-based instrument that is simple to administer and produces a reliable estimate of patient standard gamble (SG) utilities for current health status.
Methods: A 1-page paper questionnaire instrument, paper standard gamble (PSG), was designed to estimate SG utilities. We performed two studies to assess the validity of PSG. First we compared PSG and SG utilities for current health in patients with prostate cancer. They randomly received either PSG followed by SG or vice versa, always with an intervening SF-12. In the second validity study, we assessed the test-retest reliability of PSG by administering it to prostate cancer patients twice, at least 2 weeks apart.
Results: In the first study, utilities were assessed in 64 men (32 per SG/PSG order group). A paired-comparison *t* test suggested no difference between SG and PSG (mean difference = −0.007; 95% confidence interval (CI), −0.022 to 0.008). The concordance correlation coefficient was 0.92 (95% CI, 0.79 to 0.99). In the second study, test and retest PSGs were available for 184 patients. The concordance correlation coefficient was 0.88 (95% CI, 0.73 to 0.94).
Conclusions: These data suggest that PSG may serve as a reliable substitute for SG when current health utility is of interest. PSG may have particular advantages for acquisition of health-related quality-of-life data in longitudinal studies.

Keywords: Utilities, Standard gamble, Quality of life, Prostate cancer

When comparing groups of patients, such as those participating in treatment arms of a randomized clinical trial, quality of life is often important. Unfortunately, traditional quality-of-life questionnaires may lead to conflicting results. For example, treatment A may prolong life but inflict a lower quality of life than treatment B, forcing the patient

The authors thank Scott B. Cantor, PhD, for his review of this manuscript. This study was funded in part by the Leon Lowenstein foundation.

https://doi.org/10.1017/S0266462303000138 Published online by Cambridge University Press

and physician to trade off one for the other when deciding on a treatment. The field of oncology provides several examples of this phenomenon, where patients willingly accept a treatment strategy with shorter life expectancy in anticipation of a higher quality of life (8;18;22;34;35;36;45;50;54). Another area of conflict results from the multiple endpoints measured in most quality-of-life questionnaires (17). If some quality-of-life endpoints are better with one particular treatment and other quality of life endpoints are worse, it becomes difficult to judge whether the particular treatment is superior overall. For treatment decision making, these multiple quality-of-life criteria must be weighed and compared with survival benefits to select a single treatment strategy (17).

Utility assessment is an attractive health-related quality-of-life measure because it produces a single overall quality-of-life value. In theory, it incorporates all aspects of health into a single measure on a scale with endpoints of "death" and "perfect health." Such a scale is appropriate to weight survival times for the calculation of quality-adjusted life-years (QALYs), the dominant measure of health improvement for decision analyses and cost-effectiveness analyses (56). With QALYs, quantity and quality of life are simultaneously considered when comparing alternative treatment strategies. As an outcome measure, quality-adjusted survival could be calculated by following a patient for life, continually assessing his current health utility, and integrating this function (12;16;17;20;43).

With the standard gamble (SG) utility assessment method, patients are offered an option (e.g., a hypothetical pill) that will confer upon them either perfect health or death (4). The risk of death is titrated to find the maximum that the patient is willing to endure for a chance at obtaining perfect health. The reliability of the SG is generally very good (1;15;49), and patients generally understand it (2). Because it involves uncertainty, a characteristic of practically all medical decisions (2), it is a true utility assessment method (2;26;41). It is the only method consistent with the von Neumann and Morgenstern axioms of decision theory (10;21;31;40), and the only utility measure for which expected value is meaningful (55). While SG may not be the perfect utility measure for all occasions, it is certainly desirable in many contexts.

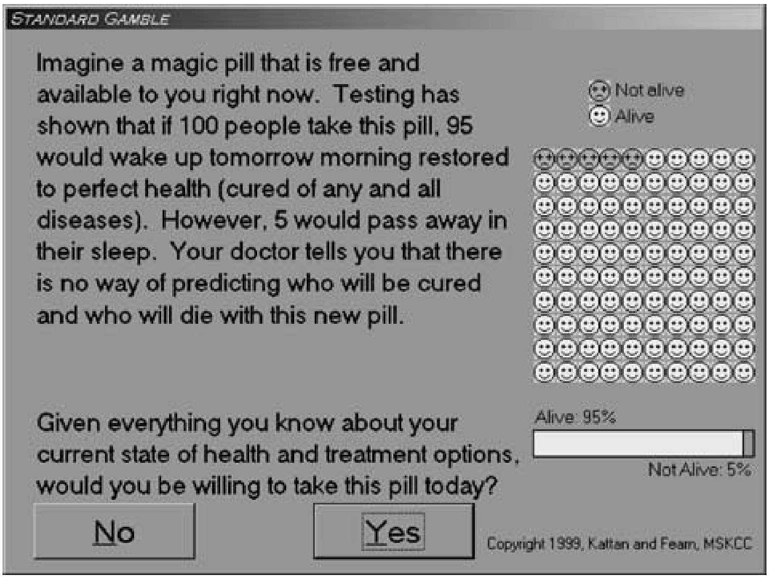
A downside to the SG method is the expense associated with its measurement (46). Most contemporary assessments involve face-to-face interaction (10) with trained interviewers or sophisticated computer software (15;28). Generally, utility assessment involves travel by either patient or facilitator. Thus, as a continually assessed measure of quality of care, the SG is infrequently used because of its complexity and cost of administration. While questionnaire-based utility assessment in general is desirable, replacing the computerized standard gamble becomes particularly attractive (16).

We sought to develop and test a paper questionnaire, called paper standard gamble (PSG), which would provide estimates similar to those obtained with the computerized SG. If successful, a paper questionnaire would be much easier to administer than the interviewer-based or computerized forms of SG (33;46). Another advantage to a paper proxy for the SG (that does not require an interviewer) is that it would greatly facilitate population-based utility assessments by mail (33). Previous studies have suggested that SG has high test-retest reliability, with an intraclass correlation coefficient of 0.77 to 0.79 (1) and a test-retest correlation coefficient of 0.85 to 0.88 (27). Indication that the concordance between PSG and SG was near the reliability of SG would support the replacement of PSG for SG when ease of administration was of concern.

METHODS

Two separate studies were conducted to assess the psychometric properties of PSG. The first study compared the PSG with SG measured in the same individual (i.e., criterion validity). The second study examined the test-retest reliability of PSG.

https://doi.org/10.1017/S0266462303000138 Published online by Cambridge University Press



STANDARD GAMBLE

Imagine a magic pill that is free and available to you right now. Testing has shown that if 100 people take this pill, 95 would wake up tomorrow morning restored to perfect health (cured of any and all diseases). However, 5 would pass away in their sleep. Your doctor tells you that there is no way of predicting who will be cured and who will die with this new pill.

Given everything you know about your current state of health and treatment options, would you be willing to take this pill today?

Figure 1. A screen capture from the standard gamble software. This was used as the criterion standard in test of PSG validity.

Criterion Validity

Our overall design was to compare the PSG as an index test with the SG by computer as a reference test. To perform the SG, we wrote software resembling that of IMPACT, developed by Lenert and colleagues (29). The interface features large fonts and various graphics to convey the standard gamble probabilities (Figure 1) (29;47). We worded the standard gamble as a "magic pill" that would either cause the patient to wake up the following morning in perfect health or cause the patient to die in his or her sleep after taking the pill (4). The interval division method (27) was used to locate the maximum risk of death the patient would be willing to accept within 1% precision. The particular interval division algorithm we used was as follows. The patient was first asked to consider a pill with 5% risk of death. If accepted, a pill with 50% risk of death was presented. If accepted, a pill with 99% risk of death was offered. If the initial pill with 5% risk of death was refused, a pill with 1% risk of death was offered. Within these boundaries, when a pill was refused, a pill was offered with risk equal to the midpoint between the refused pill and the previously accepted pill with maximum risk, rounded to the 1% risk level.

We developed a pencil-and-paper questionnaire that closely resembled the wording of the standard gamble. The risk of death was ordered from lowest to highest. Although this design is easy for the patient to follow, it theoretically lacks efficiency because it requires the patient to answer every question and it potentially creates utilities that are biased high due to its top-down order. Because branching on paper may be difficult for some patients to follow, our approach was simply to pilot test this theoretically inferior questionnaire and assess its performance and potential bias. The questionnaire appears in Appendix 1. Of notice, the granularity of PSG is skewed to achieve a finer elicitation at higher utility values. This design is intentional because previous work by us (22;24) and other investigators (8;25) demonstrated that the overwhelming majority of patients scored at the higher end of the utility spectrum. The region of fine granularity could theoretically be shifted to accommodate different patient populations. We conducted a pilot test of the instrument among methodologists for initial feedback. The typical response from them was that it resembled the standard gamble so closely that formal testing of the instrument was not

https://doi.org/10.1017/S0266462303000138 Published online by Cambridge University Press

necessary. We interpreted these comments as establishing face validity of the instrument but proceeded with formal testing because error and bias issues needed to be addressed empirically.

Experimental Design. The PSG and SG were administered to patients previously diagnosed with prostate cancer while they waited for their clinic appointment. The order of presentation was randomized (i.e., half the patients received the PSG prior to the SG, and vice versa) so that we could test for an ordering effect. It was imperative that we administer both instruments in a period of time in which the patient's opinion of his health would not change, so the instruments were administered sequentially. However, we administered an SF-12 questionnaire (51) between the two utility assessments to provide a distraction between utility assessments in an attempt to lessen the tendency for the patient to simply match responses without thinking about the utility assessment exercise. With statistical analysis, we tested whether in fact one instrument affected the response to the other (i.e., the degree to which the SF-12 was ineffective in this regard).

Prior to each active clinic day, we obtained a list of patients scheduled for a clinic visit. A research fellow approached consecutive patients in an outpatient prostate cancer clinic as they waited to see their attending physician. Patients were briefed on the reasons behind the study and asked if they were willing to participate. Informed consent was obtained in the presence of the research fellow. All patients were eligible for our study regardless of their treatment decisions or disease status. The only inclusion criterion for this study was the ability to read English. If a patient reported to the clinic multiple times during the study, he was only allowed to participate the first time we encountered him. Patients with prostate cancer were selected as a matter of convenience and because they have well-documented decision-making challenges ahead of them (8;14;22;25). With this disease, where quantity and quality trade-offs are ubiquitous at all stages of disease, utility assessment is a particularly valuable tool. The patient was informed that he would be asked to complete three brief questionnaires, two on paper and one on the computer. In appropriate sequence, one of these questionnaires or a laptop was placed in front of the patient and removed as soon as he finished that particular component, at which time the next questionnaire was placed in front of the patient until study completion.

Statistical Methods. Several statistical analyses were conducted to assess the ability of PSG to estimate SG. Initially, the Pearson correlation coefficient between the two instruments was computed and tested for significance. However, the correlation coefficient assesses only strength of correlation and not calibration. Therefore, a Bland-Altman plot (3) was constructed to visualize potential bias in the PSG across its scale. To quantify and test for an overall bias effect, a paired comparison *t* test was performed by subtracting each patient's PSG from his SG and testing whether the mean difference across patients deviated from zero. The potential for an effect of the order of administration was evaluated by comparing the SG-PSG mean differences between order groups (SG prior to PSG vs. PSG prior to SG) with a two-sample *t* test. Conformity of PSG and SG was measured with the concordance correlation coefficient (13), a measure that is very similar to the better-known intraclass correlation coefficient but with arguably better theoretical properties (13) and computer code (30) available for bootstrapping in S-Plus 2000 Professional Software (Insightful Corp, Seattle). The concordance correlation coefficient (CCC) measures the agreement between the paired PSG and SG values versus a 45° line (i.e., the line where PSG = SG in a scatterplot). Bootstrapping with 1,000 resamples was used to derive 95% confidence intervals for this coefficient. Finally, we determined whether the variance in SG utilities was affected by exposure to the PSG and SF-12, which might be expected if the PSG were a poorly comprehended instrument that only biased the subject who was trying

https://doi.org/10.1017/S0266462303000138 Published online by Cambridge University Press

to match his SG to his prior PSG. Variance comparison was accomplished with use of the folded form F statistic (42).

Sample Size Considerations. The purpose of the PSG instrument was to estimate the SG score very closely. We were particularly concerned by the threat of making a type II error—concluding that the SG and PSG utilities were equivalent when they are not. Thus, we desired a very statistically powerful sample size to guard against a type II error. Given that our primary test of the PSG instrument was a paired comparison t test, and assuming a type I error rate of 5%, we would have 98% power (and thus a 2% type II error rate) to detect a moderate effect size difference (11) between the PSG and SG with a sample size of 64 patients.

Test-Retest Reliability

Experimental Design. As part of a larger questionnaire, PSG was administered to patients previously diagnosed with prostate cancer. While in the clinic, patients were asked to complete the questionnaire. Then, upon leaving the clinic, patients were given a second copy of the questionnaire and asked to complete it in approximately 2 weeks. A self-addressed, stamped envelope was also provided. Patients were telephoned as a reminder when they were 7 days late in returning the questionnaire. Another copy was mailed when patients were 14 days late.

Statistical Methods. Test-retest reliability was quantified with the concordance correlation coefficient. Bootstrapping with 1,000 resamples was used to derive 95% confidence intervals (CI).

RESULTS

Criterion Validity

Utilities were assessed in 64 patients; 32 had their PSG assessed prior to the SG, and 32 were assessed in reverse order, with all patients completing the SF-12 between utility assessments. No patient refused participation or failed to complete the study. Complete assessment generally took less than 5 minutes per patient.

The PSG was scored by determining the midpoint between adjacent questions where the patient responded "Yes" and "No" (Appendix 1). No patient responded inconsistently (i.e., answering "Yes" to a pill that conveyed more risk of death than a pill previously refused, or vice versa). Of the 64 patients, 53 circled every question on the PSG, as we had envisioned. Three patients circled multiple "Yes" responses at the top of the instrument and left the remaining lines blank. Of note, these patients were not interrupted or pulled away for a meeting with their doctor. One of us (MWK), who was blinded to the patients' SG responses, scored these as if the uncircled questions were answered with "No." Eight patients circled one or more "Yes" responses at the top of the instrument, followed by a single "No," and then left the remaining lines blank. These PSG scores were computed as usual: the midpoint between the adjacent "Yes" and "No" responses. A scatterplot of the PSGs versus the SGs (both jittered to reveal duplicate points) appears in Figure 2. The ideal reference line is overlaid. The Pearson correlation between the PSG and SG was 0.94. The Bland-Altman plot appears in Figure 3. The horizontal line, which represents the SG-PSG mean, is very near 0, suggesting good agreement overall. By paired comparison t test, this mean difference (−0.007) was not significantly different from 0 ($p = .34$; 95% CI, –0.022 to 0.008). Patients with very high utilities (>0.9) may tend to have higher SGs than PSGs, with the reverse trend for patients in the middle of the spectrum. To investigate this potential nonlinear trend further, we fit a flexible linear regression of SG on PSG using a restricted

https://doi.org/10.1017/S0266462303000138 Published online by Cambridge University Press

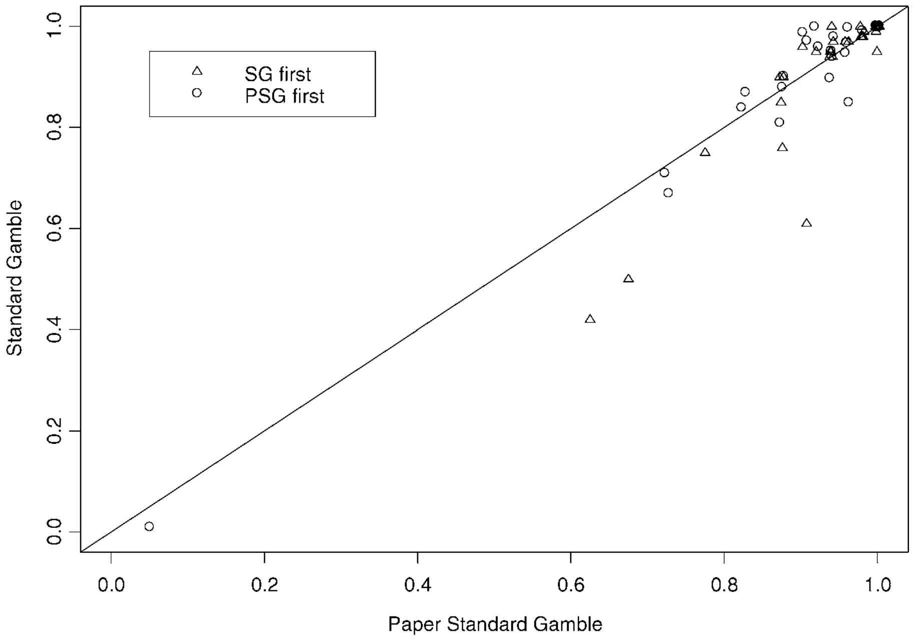


Figure 2. Scatterplot of standard gamble vs. paper standard gamble. Triangles represent subjects who completed the SG prior to the PSG, and circles represent subjects who completed the PSG prior to the SG. The line represents the ideal reference line.

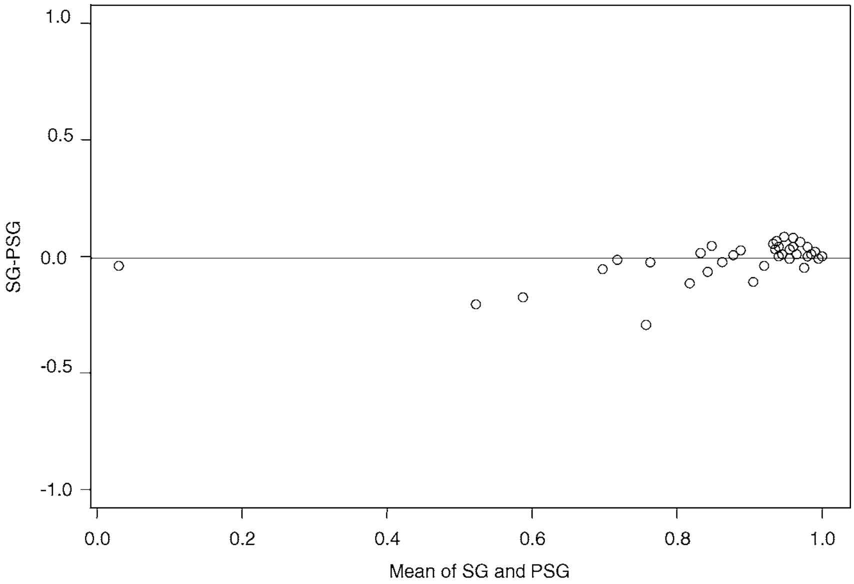


Figure 3. Bland-Altman plot. Circles represent patients in the study. The vertical axis is the patient's PSG subtracted from his SG. The horizontal axis is the patient's mean utility [(PSG + SG)/2]. The line indicates the mean difference (SG − PSG) for the 64 patients.

https://doi.org/10.1017/S0266462303000138 Published online by Cambridge University Press

cubic spline to allow for a potentially nonlinear effect. The nonlinear spline did not improve model fit ($p = .93$), providing no support for a nonlinear trend in the relationship. The concordance correlation coefficient was high (CCC = 0.92; 95% CI, 0.79 to 0.99).

The difference in utilities was further examined to assess whether the order of utility assessment was associated with a change in the difference between the PSG and SG scores. The mean difference in the group that had PSG assessed prior to SG was not different from the mean difference in the group that had PSG assessed after SG ("PSG prior" mean difference minus "PSG after" mean difference = 0.015;95% CI, –0.007 to 0.053, $p = .13$). The folded form F statistic was used to test for inequality in variances of the SG when administered before versus after the PSG, and no evidence of a difference was found ($p = .20$).

We also examined whether the SG estimates for patients who received SG first differed from the PSG estimates from patients who received the PSG first. A difference in these means would imply a pure instrument effect, since each is occurring first in the sequence of instruments. By two-sample *t* test, no difference was observed ($p = .91$; 95% CI for difference, −0.09 to 0.08).

Test-Retest Reliability

Between October 17, 2000 and July 17, 2001, we enrolled 439 patients in a health-related quality of life protocol for prostate cancer patients. Of these, 221 returned a second copy at least 14 days after the first questionnaire. A scatterplot of the test and retest PSG (both jittered to reveal duplicate points) appears in Figure 4. The ideal reference line is overlaid. The concordance correlation coefficient was 0.88 (95% CI, 0.73 to 0.94).

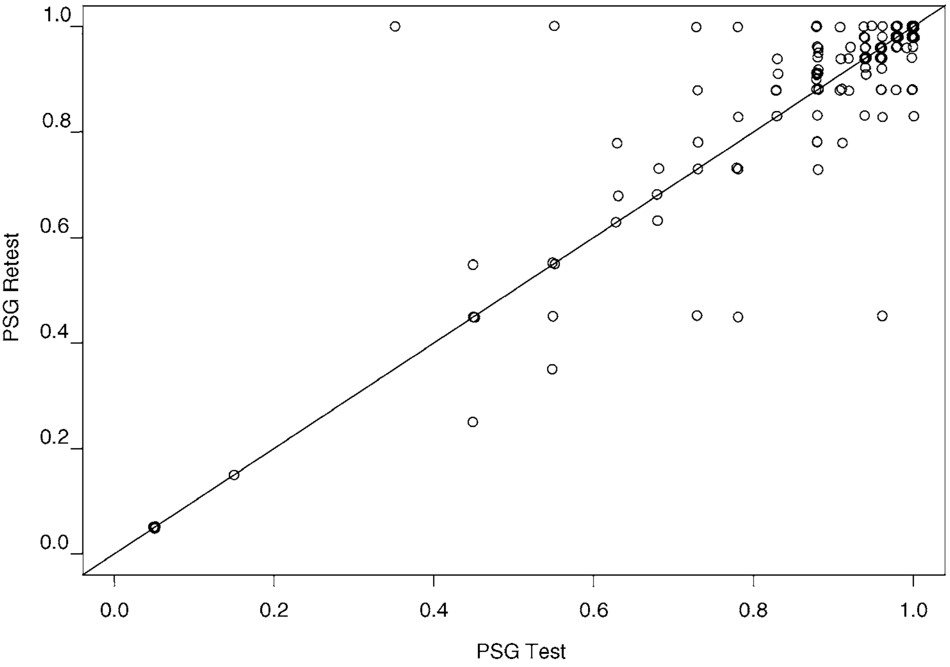


Figure 4. Scatterplot of PSG retest vs. PSG test. Circles represent individual patients; the line represents the ideal reference line.

https://doi.org/10.1017/S0266462303000138 Published online by Cambridge University Press

DISCUSSION

Previous attempts to measure utilities by paper in individual patients have been unsatisfactory. The Health Utilities Index (HUI), versions II and III, are paper questionnaires that produce SG estimates based on conversion from rating scale preferences for hypothetical health states (5). HUI is intended for applicability to groups of patients rather than individuals. Patient responses to rating scales are mapped to SG estimates from power curves derived by the HUI authors. However, little is available in peer-reviewed literature regarding the HUI-SG agreement. Given that the SG involves individual risk tolerance, it seems unlikely that a population-based health index could accurately translate rating scale scores into SG scores across individuals whose risk tolerances likely differ at the individual level. In fact, it is common to find patients in the same clinical health state who have very different utilities and degree of bother (37). An additional limitation of the Health Utilities Index instrumentation is that it is substantially longer than the 1-page PSG instrument.

Some investigators have attempted to predict the SG from an SF-36 questionnaire (52), a highly attractive solution to paper-based utility assessment. Bryce et al. (7) found the SF-36 explained only 20% of the variance in the SG. Bosch and Hunink (4) similarly found that only 14% of the variance in the SG was explained by the SF-36. Brazier et al. (6) were considerably more successful using a more elaborate model, explaining 50% of the variance in the SG with the SF-36 (6). However, our data show that the PSG explains 88% of the variance in the SG. In general, it would seem unlikely that any health status index could accurately predict the SG since the list of potential health conditions bothering the patient is practically unlimited. A health status measure designed to predict utility (53) would likely omit some aspect of health, suggesting that a paper-based measure of utility may have to be a global utility measure.

An interesting difficulty with generalizing our results to other utility assessment reports is that the scenario we asked patients to consider was somewhat unique. We wanted patients not to focus solely on their current health at that instant, but instead on their perception of their health trajectory. The difference can be critical and especially relevant to prostate cancer, where a man is typically diagnosed without symptoms but who now has fear of his future health. He may be happy with his current health as long as he can maintain it, but what worries him is the degeneration, particularly if the disease is left untreated. In general, if a patient is deeply concerned with what lies ahead, utility assessment should reflect this concern, relative to a similar patient without present complaints and with no future concerns. These two patients have different perceptions of their health and should have different utilities. We therefore applied standard gambles that encouraged patients to consider their health trajectory, along the lines of modifications we suggested for adapting the time trade-off utility assessment method to also reflect degenerative or life-threatening conditions (23).

A limitation of our criterion validity study may be the very brief time between administering the PSG and SG. In our experimental design, an SF-12 was sandwiched between the PSG and SG, whose order was randomized. We obtained very high correlation between SG and PSG estimates, which could be related to the short time between administrations. Were we testing the reliability of the SG, whereby the SG was administered both times, such a short interval would clearly be cause for concern because this would likely inflate the reliability estimation (9). However, these were two different instruments whose look and feel are distinctly different: the SG is a highly graphical computer experience, while the PSG is a text-based paper questionnaire. If PSG scores were, in reality, not so closely related to SG scores, and SG scores were reliable and understood by patients, one would not expect to obtain data similar to ours. If patients did not understand the PSG, but thought our experiment was a game whereby they were supposed to match their PSG and SG estimates, the SG scores would have been expected to be distributed differently, depending on whether

https://doi.org/10.1017/S0266462303000138 Published online by Cambridge University Press

they preceded or followed the PSG. The two SG distributions were very similar ($p = .20$). Thus, patients would have to have thought that the PSG and SG questions were so similar that they should, without thinking, provide the same answers. We would view that more as a strength of the PSG instrument rather than a weakness of the design. Because the SG is potentially, and appropriately, affected by many aspects and feelings toward one's health, even a visit with the doctor has the potential to affect the SG (19). Therefore, we had to assess the patients' utilities by both methods in a very short period of time (e.g., prior to meeting with the physician) to ensure that the patient's opinion about his health had not changed. Especially in the clinical environment where we tested the PSG, we would not have expected a patient's utility to measure the same before and after the encounter with the physician (an interesting question for future research).

Another limitation of this study is the relative homogeneity of subjects. As prostate cancer patients at a tertiary care institution, these were mostly white men with a median age of 62 years and above average income and education. Generalization to other patient groups should be studied, especially in a more diverse patient population with a wider distribution of expected utilities. Moreover, further research should address the test-retest reliability and other psychometric properties of PSG to investigate whether patients understand the PSG concept.

A third limitation, regarding the questionnaire rather than the design *per se*, is that the PSG uses top-down titration rather than the interval division method, which is considered to be less biased. However, we found no evidence of an empirical bias. On average, PSG scores were not higher than SG scores. Previous studies have found that top-down titration produces less variable but higher utilities than does bisecting titration (27), possibly due to fatigue associated with the large number of questions asked of the patient with the bisecting method. Percy and Llewellyn-Thomas (39) found that top-down titration scores differed from bottom-up scores. However, it is unknown which, if either, would have significantly differed from the theoretically preferred method of interval division. Fatigue from top-down titration may not be an issue with the PSG because it is a short instrument.

A troubling aspect in utility assessment is the lack of a gold standard, even within SG utility assessment. We chose to develop our own software and base it largely on the extensive work of Lenert (28;29). We conveyed probabilities as happy and sad faces along with a bar chart. Numerous design options are available in this setting, including whether to segregate or randomize the sad faces, which colors to use, and the precise choice of wording. The lack of a proven gold standard makes these choices rather arbitrary until one configuration is demonstrated to be superior to others. Therefore, our agreement between PSG and SG may change with an alternative format of the SG, and this aspect should be studied. Because we are measuring current health utility rather than utility of a particular hypothetical health condition, we chose not to provide the face and voice of someone describing their health conditions (10) that may or may not be relevant to the particular individual being assessed, an approach which may be very valuable for hypothetical health states.

It is also unknown whether SG, as a reference test, should be facilitated by an interviewer. The argument for facilitation is that a trained facilitator helps the patient understand what is being asked during utility assessment. The argument against facilitation is that the trained facilitator may bias the respondent into utilities that the facilitator thinks are reasonable. Given this uncertainty as to which design is preferable, and the particular question in this study of whether two assessment methods yield equivalent utilities, we felt it was more appropriate that the SG utility not be facilitated. Thus, our facilitator did not help the patient with either the PSG or SG assessments. When the patient asked a content-related question, the facilitator was instructed to reply with "do the best you can; there are no right or wrong answers." His only responsibilities were obtaining informed consent, coordinating interviews, and resolving technical issues such as computer (or pencil) malfunctions.

https://doi.org/10.1017/S0266462303000138 Published online by Cambridge University Press

The lack of a gold standard in utility assessment also hampers the utilization of the SF-12 responses. We did not analyze or report these data because the SF-12 does not measure all aspects of current health, which is an infinite list of mental, social, physical, and other conditions. Therefore, correlation between PSG and SF-12 domains is of little interest and would neither validate nor nullify PSG, especially given the poor correlation between SG and the SF-12 (44).

Though additional psychometric work remains, the PSG appears to be a promising questionnaire for utility assessment. If validated in future studies, the PSG has the potential to contribute at the basic science level to our understanding of utilities. Many questions remain as to the stability of utilities over time (32) and relationships with various patient and process factors (48). PSG would seem to facilitate repeat current health utility assessment over large samples of patients and make this data collection much more feasible. While computerized utility assessment is attractive for other reasons (e.g., automated data entry, multimedia and personalized presentations for hypothetical health states), paper-based assessment for current health utility is substantially less expensive to conduct across broad and diverse populations.

POLICY IMPLICATIONS

The PSG may add benefit to the conduct of clinical trials where a comprehensive (i.e., quality and quantity of life) judgment is desired. Patients could, as part of their routine follow-up, be sent the PSG instrument, and their scores could be integrated in quality-adjusted survival computations. Use of the PSG in observational (i.e., nonrandomized) studies may also prove useful, but additional care is needed to consider the potential confounding issue that baseline PSG scores may in fact be directing the patient's treatment choices. For example, patients with lower utilities (and lower PSG scores) may be more likely to choose more aggressive therapies, which involve greater risks and may tend to result in yet lower utilities (and thus lower PSG scores). A baseline PSG assessment could be used in the observational setting to help correct for this confounding, by allowing adjustment for the baseline differences in utility. This could be accomplished through multivariable analyses or computation of change scores (e.g., follow-up minus baseline PSG).

Use of PSG might also facilitate economic analyses, which utilize dollars per quality-adjusted life year as the metric of cost-effectiveness. Here, the PSG could be utilized in the cost-effectiveness analysis to help measure the denominator through prospective assessment of patients in each screening or treatment strategy in the analysis. Furthermore, it seems possible to adapt the PSG to the prospective medical decision-making setting, whereby a patient was presented with condition-specific PSGs (rather than the current health PSG reported here). These hypothetical health-state PSGs could then be entered into a decision analysis to support real-time patient-specific medical decision making (38). However, future research is needed to demonstrate this feasibility.

In summary, these data suggest that a 1-page paper questionnaire is a reliable measure of patient utility. We found that the PSG accurately reflected the SG in 64 men, and that the PSG had excellent test-retest reliability in a sample of 184 men. The PSG potentially can play a role in various analyses of screening or treatment options involving economic concerns or where quality-of-life concerns have the potential to override survival benefits. With further psychometric evaluation, PSG may prove useful in an environment in which a traditional SG utility is desired.

REFERENCES

1. Bakker C, Rutten-van Molken M, Hidding A, et al. Patient utilities in ankylosing spondylitis and the association with other outcome measures. *J Rheumatol.* 1994;21:1298-1304.

https://doi.org/10.1017/S0266462303000138 Published online by Cambridge University Press

2. Bennett KJ, Torrance GW. *Measuring health state preferences and utilities: Rating scale, time trade-off, and standard gamble techniques—Quality of life and pharmacoeconomics in clinical trials.* 2nd ed. Philadelphia: Lippincott-Raven Publishers; 1996:253-265.
3. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet.* 1986;1:307–310.
4. Bosch JL, Hunink MG. The relationship between descriptive and valuational quality-of-life measures in patients with intermittent claudication. *Med Decis Making.* 1996;16:217-225.
5. Boyle MH, Furlong W, Feeny D, et al. Reliability of the health utilities index–mark iii used in the 1991 cycle 6 Canadian General Social Survey health questionnaire. *Qual Life Res.* 1995;4:249-257.
6. Brazier J, Usherwood T, Harper R, Thomas K. Deriving a preference-based single index from the UK SF-36 health survey. *J Clin Epidemiol.* 1998;51:1115-1128.
7. Bryce C, Angus D, Stahl J, et al. Using health status measures to predict utilities in patients with end-stage liver disease. *Med Decis Making.* 1999;19:534.
8. Cantor SB, Spann SJ, Volk RJ, et al. Prostate cancer screening: A decision analysis. *J Fam Pract.* 1995;41:33-41.
9. Churchill DN, Torrance GW, Taylor DW, et al. Measurement of quality of life in end-stage renal disease: The time trade-off approach. *Clin Invest Med.* 1987;10:14-20.
10. Clarke AE, Goldstein MK, Michelson D, et al. The effect of assessment method and respondent population on utilities elicited for Gaucher disease. *Qual Life Res.* 1996;6:169-184.
11. Cohen J. *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
12. Cole BF, Gelber RD, Goldhirsch A. Cox regression models for quality adjusted survival analysis. *Stat Med.* 1993;12:975-987.
13. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures: Statistics and strategies for evaluation. *Control Clin Trials.* 1991;12(suppl):142S-158S.
14. Fleming C, Wasson JH, Albertson PC, et al. A decision analysis of alternative treatment strategies for clinically localized prostate cancer. *JAMA.* 1993;269:2650-2658.
15. Froberg DC, Kane RL. Methodology for measuring health-state preferences, ii: Scaling methods. *J Clin Epidemiol.* 1989;42:459-471.
16. Glasziou P, Cole BF, Gelber RD, et al. Quality adjusted survival analysis with repeated quality of life measures. *Stat Med.* 1998;17:1215-1229.
17. Glasziou P, Simes R, Gelber R. Quality adjusted survival analysis. *Stat Med.* 1990;9:1259-1276.
18. Grann VR, Panageas KS, Whang W, et al. Decision analysis of prophylactic mastectomy oophorectomy in brca1-positive or brca2-positive patients. *J Clin Oncol.* 1998;1:979-985.
19. Hilden J, Glasziou P. Regret graphs, diagnostic uncertainty and Youden's index. *Stat Med.* 1996;15:969-986.
20. Hwang JS, Tsauo JY, Wang JD. Estimation of expected quality adjusted survival by cross-sectional survey. *Stat Med.* 1996;15:93-102.
21. Kaplan RM, Feeny D, Revicki DA. Methods for assessing relative importance in preference based outcome measures. *Qual Life Res.* 1993;2:467-475.
22. Kattan MW, Cowen ME, Miles BJ. A decision analysis for treatment of clinically localized prostate cancer. *J Gen Intern Med.* 1997;12:299-305.
23. Kattan MW, Fearn PA, Miles BJ. Time trade-off utility modified to accommodate degenerative and life-threatening conditions. Paper presented at the 2001 AMIA Symposium, 2001; Washington, DC.
24. Kattan MW, Hu J, Cowen ME, et al. Do easier utility assessment methods work? Comparisons of traditional approaches with phone interviews and the SF-12 [abstract]. *Med Decis Making.* 1997;17:537.
25. Krahn MD, Mahoney JE, Eckman MH, et al. Screening for prostate cancer. *JAMA.* 1994;272:773-780.
26. Lalonde L, Clarke AE, Joseph L., et al. Conventional and chained standard gambles in the assessment of coronary heart disease prevention and treatment. *Med Decis Making.* 1999;19:149-156.

https://doi.org/10.1017/S0266462303000138 Published online by Cambridge University Press

27. Lenert LA, Cher DJ, Goldstein MK, et al. The effect of search procedures on utility elicitations. *Med Decis Making*. 1998;18:76-83.
28. Lenert LA, Hornberger JC. Computer-assisted quality of life assessment for clinical trials. *Proc JAMIA Annu Fall Symp*. 1996:992-996.
29. Lenert LA, Michelson D, Flowers C, Bergen MR. Impact: An object-oriented graphical environment for construction of multimedia patient interviewing software. Paper presented at the 19th Annual Symposium on Computer Application in Medical Care: A Conference of the American Medical Informatics Association, Oct. 28–Nov. 1, 1995, New Orleans, La.
30. Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*. 1989;45:255-268.
31. Llewellyn-Thomas H, Sutherland HJ, Tibshirani R, et al. Describing health states: Methodologic issues in obtaining values for health states. *Med Care*. 1984;22:543-552.
32. Llewellyn-Thomas HA, Sutherland HJ, Thiel EC. Do patients' evaluations of a future health state change when they actually enter that state? *Med Care*. 1993;31:1002-1012.
33. Lundberg L, Magnus J, Isacson DG, Borquist L. The relationship between health-state utilities and the SF-12 in a general population. *Med Decis Making*. 1999;19:128-140.
34. McNeil BJ, Weichselbaum R, Pauker SG. Fallacy of the five-year survival in lung cancer. *N Engl J Med*. 1978;299:1397-1401.
35. McNeil BJ, Weichselbaum R, Pauker SG. Speech and survival: Tradeoffs between quality and quantity of life in laryngeal cancer. *N Engl J Med*. 1981;305:982-987.
36. McQuellon RP, Muss HB, Hoffman SL, et al. Patient preferences for treatment of metastatic breast cancer: A study of women with early-stage breast cancer. *J Clin Oncol*. 1995;13:858-868.
37. Nease RFJ, Kneeland T, O'Connor GT, et al. Variation in patient utilities for outcomes of the management of chronic stable angina. *JAMA*. 1995;273:1185-1190.
38. Pauker SG. Coronary artery surgery: The use of decision analysis. *Ann Intern Med*. 1976;85:8-18.
39. Percy ME, Llewellyn-Thomas H. Assessing preferences about the DNR order: Does it depend on how you ask? *Med Decis Making*. 1995;15:209-216.
40. Read JL, Quinn RJ, Berwick DM, et al. Preferences for health outcomes: Comparison of assessment methods. *Med Decis Making*. 1984;4:315-329.
41. Revicki DA, Kaplan RM. Relationship between psychometric and utility-based approaches to the measurement of health-related quality of life. *Qual Life Res*. 1993;2:477-487.
42. SAS Institute I. The folded form *f* statistic. *Sas/stat user's guide,* version 6. 4th ed. Cary, NC: SAS Institute Inc; 1989:1636.
43. Shen L, Pulkstenis E, Hoseyni M. Estimation of mean quality adjusted survival time. *Stat Med*. 1999;18:1541-1554.
44. Sherbourne CD, Sturm R, Wells KB. What outcomes matter to patients? (see comments). *J Gen Intern Med*. 1999;14:357-363.
45. Silvestri G, Pritchard R, Welch H. Preferences for chemotherapy in patients with advanced non-small cell lung cancer: Descriptive study based on scripted interviews. *BMJ*. 1998;317:771-775.
46. Stiggelbout AM, Eijkemans MJ, DeHaes HJ, et al. The "utility" of the visual analog scale in medical decision making and technology assessment. Is it an alternative to the time trade-off? *Int J Technol Assess Health Care*. 1996;12:291-298.
47. Sumner W, Nease R, Littenberg B. U-titer: A utility assessment tool. *Proc Annu Symp Comput Appl Med Care*. 1991:701-705.
48. Sylvia J, Jansen M, Stiggelbout A, et al. A shift in valuation or an effect of the elicitation procedure? *Med Decis Making*. 2000;20:62-71.
49. Torrance GW. Utility approach to measuring health-related quality of life. *J Chron Dis*. 1987;40:593-600.
50. van Roosmalen M, Verhoef L, Stalmeier P, van Daal W. Extension of a prognostic Markov model for brcai-positive women. *Med Decis Making*. 1999;19:521.
51. Ware J, Kosinski M, Keller SD. A 12-item short-form health survey: Construction of scales and preliminary test of reliability and validity. *Med Care*. 1996;34:220-233.
52. Ware JE, Snow KK, Kosinski M, Gandek B. Sf-36 health survey manual and interpretation guide. Boston, MA: New England Medical Center, The Health Institute; 1993.

https://doi.org/10.1017/S0266462303000138 Published online by Cambridge University Press

53. Weeks J, O'Leary J, Fairclough D, et al. The "q-tility index": A new tool for assessing health-related quality of life and utilities in clinical trials and clinical practice. *Proceedings of ASCO*. 1994;13:436.
54. Weeks JC, Cook EF, O'Day SJ, et al. Relationship between cancer patients' predictions of prognosis and their treatment preferences. *JAMA*. 1998;279:1709-1714.
55. Weinstein MC, Fineberg HV. *Clinical decision analysis*. Philadelphia: WB Saunders Co; 1980.
56. Weinstein MC, Siegel JE, Gold MR, et al. Recommendations of the Panel on Cost-effectiveness in Health and Medicine. *JAMA*. 1996;276:1253-1258.

APPENDIX 1

Imagine a new (make-believe) pill is now available for ***all*** your health problems. Your doctor advises you that if you take the pill today and it works, it cures every health problem you ***currently*** have for the rest of your life. However, if you take the pill today and it ***does not*** work, it causes a sudden and painless death in your sleep tonight. Your doctor has no way of predicting which patients will be cured by this new (make-believe) pill, and will support whatever decision you make. Given everything you know about your current health, how it may change in the future, and your treatment options, we want to know what you think about this pill.

→***Would you take this pill right now if you knew***. . . (Please circle "Yes" or "No" for every question.)

. . . it had a **100%** chance of cure and a **0%** risk of causing death in your sleep tonight? ***Yes*** ***No***

. . . it had a **99%** chance of cure and a **1%** risk of causing death in your sleep tonight? ***Yes*** ***No***

. . . it had a **97%** chance of cure and a **3%** risk of causing death in your sleep tonight? ***Yes*** ***No***

. . . it had a **95%** chance of cure and a **5%** risk of causing death in your sleep tonight? ***Yes*** ***No***

. . . it had a **93%** chance of cure and a **7%** risk of causing death in your sleep tonight? ***Yes*** ***No***

. . . it had a **91%** chance of cure and a **9%** risk of causing death in your sleep tonight? ***Yes*** ***No***

. . . it had a **90%** chance of cure and a **10%** risk of causing death in your sleep tonight? ***Yes*** ***No***

. . . it had a **85%** chance of cure and a **15%** risk of causing death in your sleep tonight? ***Yes*** ***No***

. . . it had a **80%** chance of cure and a **20%** risk of causing death in your sleep tonight? ***Yes*** ***No***

. . . it had a **75%** chance of cure and a **25%** risk of causing death in your sleep tonight? ***Yes*** ***No***

. . . it had a **70%** chance of cure and a **30%** risk of causing death in your sleep tonight? ***Yes*** ***No***

. . . it had a **65%** chance of cure and a **35%** risk of causing death in your sleep tonight? ***Yes*** ***No***

. . . it had a **60%** chance of cure and a **40%** risk of causing death in your sleep tonight? ***Yes*** ***No***

. . . it had a **50%** chance of cure and a **50%** risk of causing death in your sleep tonight? ***Yes*** ***No***

. . . it had a **40%** chance of cure and a **60%** risk of causing death in your sleep tonight? ***Yes*** ***No***

. . . it had a **30%** chance of cure and a **70%** risk of causing death in your sleep tonight? ***Yes*** ***No***

. . . it had a **20%** chance of cure and a **80%** risk of causing death in your sleep tonight? ***Yes*** ***No***

. . . it had a **10%** chance of cure and a **90%** risk of causing death in your sleep tonight? ***Yes*** ***No***

https://doi.org/10.1017/S0266462303000138 Published online by Cambridge University Press