

Effect of Assessment Method on the Discrepancy between Judgments of Health Disorders People Have and Do Not Have: A Web Study

Jonathan Baron, David A. Asch, Angela Fagerlin, Christopher Jepson, George Loewenstein, Jason Riis, Margaret G. Stineman, Peter A. Ubel **PROVIDE DEGREES?**

Three experiments on the World Wide Web asked subjects to rate the severity of common health disorders such as acne or arthritis. People who had a disorder (“Haves”) tended to rate it as less severe than people who did not have it (“Not-haves”). Two explanations of this Have versus Not-have discrepancy were rejected. By one account, people change their reference point when they rate a disorder that they have. More precise reference points would, on this account, reduce the discrepancy, but, if anything, the discrepancy was larger. By another account, people who do not have the disorder fo-

cus on attributes that are most affected by it, and the discrepancy should decrease when people make ratings on several attributes. Again, if anything, the discrepancy increased when ratings were on separate attributes (combined by a weighted average). The discrepancy varied in size and direction across disorders. Subjects also thought that they would be less affected than others. **Key words: PLEASE PROVIDE 4-5 KEYWORDS. (Med Decis Making 200?;XX(X):XXX-XXX)**

INTRODUCTION

Brickman and his colleagues argued that people adapt to their lives so that both the joy of positive changes and the despair of negative changes wear off over time.^{1,2} Paraplegics, after a few years, are almost as “happy” as other people. Consistent with these observations, patients with chronic health disorders sometimes do not see their disorder as very bad, compared to judgments offered by those who have not experienced the disorder.^{3,4} In these studies, patients report that their quality of life is significantly better than the public estimates it would be. For example, Sackett and Torrance⁴ found that the general public estimates the health-related quality of life (HRQoL) of dialysis at a value of 0.39 (on a scale from 0 for conditions as bad as death to 1 for perfect health), whereas dialysis patients estimate their HRQoL at 0.56. Boyd and others³ found that patients without colostomies estimate the HRQoL of living with a colostomy at 0.80, whereas patients with colostomies rate their own HRQoL at 0.92. De Wit and others provide additional evidence for dialysis patients and review earlier literature.⁵

Other studies, however, fail to find such discrepancies (e.g., Llewellyn-Thomas and others⁶ and O’Connor and others⁷). These studies often examine short-term conditions, such as the experience of radiation therapy, evaluated by the same patients, before and during the experience.

The question we address here is whether the discrepancy between the ratings of people who have and

Received DATE? from the Department of Psychology, University of Pennsylvania, Philadelphia (JB); Leonard Davis Institute of Health Economics, University of Pennsylvania (JB, DAA, MGS); Division of General Internal Medicine, University of Pennsylvania School of Medicine, Philadelphia (DAA, CJ); Center for Health Equity Research and Promotion, Philadelphia Veterans Affairs Medical Center (DAA); Veterans Administration Health Services Research and Development Center of Excellence, Veterans Affairs Ann Arbor Healthcare System and the Division of General Internal Medicine, University of Michigan, Program for Improving Health Care Decisions, Ann Arbor, Michigan (AF, PAU); Department of Social and Decision Sciences, Carnegie-Mellon University (GL); Department of Psychology, University of Michigan, Ann Arbor (JR); Department of Rehabilitation Medicine, University of Pennsylvania School of Medicine, Philadelphia (MGS). Revision accepted for publication DATE?. This study was funded by National Institutes of Health grant AG16258 (P. Ubel, principal investigator).

Address correspondence and reprint requests to Jonathan Baron, Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104-6196; e-mail: baron@psych.upenn.edu.

DOI: 10.1177/0272989X03257277

do not have a disorder—the Have–Not-have discrepancy—can be found for common disorders in a broad sample of people. We ask this by using questionnaires on the World Wide Web, available to anyone. Although our sample is surely not representative of all human beings (the population of interest), it is diverse, and it includes many people with the health disorders of interest. If the discrepancy can be demonstrated by this sort of method, then future research on it is easier than would otherwise be the case. It is important to understand the causes of this discrepancy, at least for the purpose of deciding whose judgments to use as the basis of cost-effectiveness analysis: Haves or Not-haves. We shall discuss later the arguments on both sides of this issue.

We test here two main explanations of the Have–Not-have discrepancy. In one, the meaning of the response scale changes when people evaluate disorders that they have. A term such as *good health*, which might be used for the top of a rating scale, can mean one thing to a normal healthy person—being able to play tennis or ride a bike, for example—and quite another thing to a person who has just lost a leg, where it might mean being able to go back to the office. This can happen either because people compare themselves to others with similar disorders or because they adopt different aspiration levels. In either case, the discrepancy should be reduced by making the response scale more explicit, so that the terms used to name the ends of it are not so subject to variable interpretation.

This sort of scale recalibration could also explain response shift.⁸ People undergoing cancer therapy, in some studies, retrospectively evaluate their pretherapy condition as better than they rated that condition at the time they experienced it,⁹ or as worse, if their condition improved.^{10,11} Analogous results are found for transplant patients, who rate their pretransplant quality of life (QOL) as lower after the transplant than they did at the time.^{12,13} Jansen and others,¹⁴ however, found no evidence for a response shift.

Here, in experiments 1 and 2, we test this scale-recalibration explanation by assessing the Have–Not-have discrepancy with different kinds of response scales, which differ in the clarity of their endpoint. If a scale is defined so that endpoints have a constant meaning, and the discrepancy is reduced, then the use of undefined scales is part of the explanation of the discrepancy.

The other main explanation holds that the discrepancy is the result of a focusing illusion. When imagining the situation of others who have a disorder, people focus disproportionately on what is affected by the disorder while ignoring those things that are not affected.

In the original demonstration of this focusing illusion, students in both California and the Midwest predict that they would be significantly happier living in California than in the Midwest.¹⁵ This prediction correlates strongly with how important they feel weather is to their QOL. Yet no significant difference in happiness is found between these 2 groups of students, suggesting that they focus disproportionately on the effect of weather on their QOL when they compare themselves to those in a different climate.

In an earlier study, Ubel and others¹⁶ explored whether a focusing illusion contributes to general public estimates of the QOL associated with disabilities. Subjects estimated the QOL of either paraplegia, below-the-knee amputation, or partial blindness. Then the experimenters attempted to “defocus” subjects by having them reflect on the impact of these disabilities on a wide range of life domains. This defocusing task should keep people from thinking too narrowly about the life domains affected by the disability. For example, subjects were asked to think about how such a disability would affect their family life, assuming that for most, it would have little effect. This manipulation showed no effect. The focusing hypothesis was not supported.

Ubel and others¹⁷ replicated this negative result with three other focusing tasks. One defocusing task involved asking people to imagine how the disability in question would affect 8 concrete life events, such as “paying bills and taxes” and “reading or watching TV or movies.” Another defocusing task asked people to list those things that took up the largest amount of their time on the previous day and then indicate how much the disability in question would affect those things. A 3rd defocusing task asked people to specifically think not only about things that would be made worse about the disability in question but also things that would be unchanged and things that would be better because of the disability. None of the tasks affected the magnitude of the discrepancy, measured after doing the task.

Here, we test the focusing illusion in a different way, in experiment 3, by asking whether the Have–Not-have discrepancy is reduced when subjects provide ratings of the disorder attribute by attribute. This method ensures that subjects attend to all the attributes we provide. It should thus prevent focusing, at least to some extent, and reduce the discrepancy, if the focusing explanation is at work.

In sum, our 2 main questions are whether the Have–Not-have discrepancy can be reduced by the use of well-defined scales (experiments 1 and 2, which also examine different kinds of scales) and by the use of attribute-by-attribute ratings.

Experiments 1 and 2 also address a subsidiary question: whether people think that they can adapt to a disorder better than others can. In general, people tend to think they are “above average” on all good things.¹⁸ Many studies comparing Have and Not-have are asking the Have group (those with the disorder) to rate themselves and the Not-have group to rate others. Thus, Have–Not-have is confounded with ratings of Self and Others, so a Self-Other difference could explain the discrepancy in some studies. The Have–Not-have difference could result from a belief that “I can adapt to this better than other people can,” whether the person answering has the disorder or not. We thus ask for ratings of self with the disorder, self without it, others with, and others without. One of the two Self ratings is necessarily hypothetical. A Self-Other difference would support this explanation.

It is also possible that this Self-Other difference is found mainly in those who have the disorder. If so, Self-Other would interact with Have–Not-have.

EXPERIMENT 1

The main purpose of this experiment was to look for a discrepancy in the ratings of common disorders, in which people who do not have the disorder rate it as more severe than those who have it (Not-have v. Have). The ratings say how much disutility a disorder will create. So our general hypothesis is that Haves do not think things are as bad as Not-haves think they would be. The justification for this hypothesis is that when people have found discrepancies, they have typically been in this direction.

Note that this discrepancy need not always go in this direction. For example, anxiety, depression, and pain may be underestimated by Not-haves because the symptoms have no direct external manifestation. And indeed, in this study, one of the disorders that involves pain, migraines, showed a discrepancy in the opposite direction (Have rated worse than Not-have). Although we know of no previous findings of a reversed discrepancy of this sort, Andresen and others¹⁹ found that people who suffered pain as part of their disorder rated their pain as worse than did people who did not have the disorder, and, in a different sort of study, Todorov and Kirchner²⁰ found that proxies underreported symptoms of people with disabilities.

A 2nd question was whether a discrepancy could result from vagueness in the judgment scale. By “vagueness,” we mean that a scale could plausibly be interpreted differently as a function of having the disorder or not, for example, “extremely undesirable.” People with a health disorder might evaluate having the disorder

by comparing themselves to others with the disorder, rather than to those without it. This possibility predicts that the discrepancy would be larger when the ends of the judgment scale are vague in the sense that their interpretation could depend on the presence or absence of the disorder.

We used 3 methods of eliciting judgments of undesirability: an anchored scale, a vague scale, and a happiness scale (which was also vague). The anchored scale is anchored at the bottom by *death* and at the top by the absence of the disorder being rated. The vague scale was anchored at the bottom by *extremely undesirable* and at the top by *not undesirable at all*. The happiness scale asked about the overall effect of the disorder on happiness. Both ends were vaguely described: *greatly increased happiness* and *greatly decreased happiness*. If the discrepancy results from vagueness, then it should be smaller in the anchored scale than in the other two.

The 3rd main question concerns judgments of Self versus Others. For each scale, subjects judged for themselves and for someone else. A Self-Other difference might help to explain past findings of a Have–Not-have discrepancy.

Method

Eighty subjects completed a questionnaire on the World Wide Web. Their ages ranged from 16 to 60 years (median = 34); 27.6% were male, and 15% were students. Most subjects had completed other studies on other topics at the same site. They had originally discovered it either through search engines, links from other sites (such as those listing ways to earn money on the Web), or “word of mouth.” In general, the population of visitors to this site, although not all US residents, has about the same median education and income as the US adult population. Other than the subjects’ being Web users, their most salient characteristic is that they are mostly women. The questionnaire began as follows:

Preference for Health Conditions

The following study concerns judgments of chronic health conditions. . . .

One [type of] question concerns undesirability, which means the strength of preference for not having the condition.

[Another type of] question concerns the effect of the condition on overall happiness or unhappiness.

You will make the undesirability ratings from two points of view. One is your own point of view, if you had the condition. If you have or have had it, please pay atten-

tion to the description, and think about that rather than your own case.

The other is the point of view of the average person.

In each case, imagine that either you or the average person has had the condition for 6 months. Also, the condition will not change in the foreseeable future. It will not get better, and it will not get worse. This is important. Do not suppose that it will improve.

A typical item (1 screen) using the anchored scale for both Self and Other ratings is as follows:

Item 1 out of 45:

Suppose that you had the condition: **bad knee—running is painful.**

For yourself, on a scale of undesirability in which 0 is “not having” bad knee (with everything else the same) and 100 is “imminent death,” where would you put **bad knee**?

Now suppose that **the average person** had **bad knee (running is painful)**.

Where do you think that the average person would put **bad knee** on the same scale?

The description for the vague scale was “on a scale of undesirability in which 0 is ‘not undesirable at all’ and 100 is ‘extremely undesirable.’” The description for the happiness scale was “on a scale of happiness/unhappiness in which 5 is ‘greatly increased happiness from the condition,’ 0 is ‘no change in happiness from the condition,’ and -5 is ‘greatly decreased happiness from the condition.’” (Formatting is omitted here.) The scale types were blocked (items of each scale type presented successively) for each subject, and the order of the blocks was randomized across subjects.

With each scale type, the subject rated the following 15 health conditions, shown here with their brief descriptions:

Asthma	Attacks of breathing difficulty, 1 per week
Chronic back pain	Running is impossible and walking difficult
Bad knee	Running is painful
Inability to walk	Requires a wheelchair
Insomnia	2 h less sleep than desired on most nights
Migraines	Debilitating 2-h headaches, 2 per week
Short stature	6 inches shorter than average for gender
Excessive weight	50% more than normal weight
Nearsightedness	Glasses required
Partial deafness	Hearing aid required
Nightmares	Frightening dreams most nights
Acne	Pimples all over face

Smoking habit	Pack a day of cigarettes
Arthritis	Pain in hips or shoulders with any movement
Heart disease	Chest pain from walking or other activity

After completing all the ratings, the subjects were asked, “Which of the following conditions have you had yourself for at least a year? (You may include conditions that were more or less severe than the descriptions used here.)” They saw a list of the conditions (disorders), with a letter before each, and they were asked to type the letters.

Results

Not-Have-Have Discrepancy

Because each subject answered about several disorders, we could test the Not-have–Have discrepancy within subjects, subtracting the mean response to the disorders each subject had from the mean response to the disorders she did not have, thus gaining power by removing many of the effects of individual differences in interpretation. However, a straightforward within-subject comparison would likely yield a discrepancy simply because the less severe disorders are more frequent. To remove the effects of such differences among disorders, before doing the within-subject comparison, we standardized the ratings for each disorder across subjects (separately for Self and Other ratings with each of the 3 scales).

This standardization yielded a matrix (for each measure) with subjects as rows and disorders as columns, with some of the cells categorized as Have and most as Not-have. If a discrepancy exists, the Have cells will be rated less severe than the Not-have cells. We tested the hypothesis using subjects as the unit of analysis (mean standardized Not-have minus mean standardized Have for each subject) because, in principle, it could be done with a very small number of disorders.

Before doing statistical tests, we eliminated disorders that were rare in our sample. Because of the standardization, rare disorders could end up with extreme z scores. As it happened, 4 of the disorders (inability to walk, partial deafness, nightmares, and heart disease) occurred in 4 subjects or fewer, and all others occurred in 11 subjects or more, so we dropped these 4 health disorders from further analysis of the Not-have–Have discrepancy.

We then computed the mean of the standardized Have disorders for each subject, separately for Self and Other ratings and for the different scales and the mean

Table 1 Relationship between Scale Type and Discrepancy, in *z* Scores of Severity (mean of Self and Other, common disorders only), Experiment 1

Scale Type	Haves	Not-haves	<i>P</i>
Anchored	-0.08	0.04	0.0430
Vague	-0.03	0.02	<i>ns</i>
Happiness	-0.13	0.04	0.0479
Overall	-0.08	0.03	0.0291

of the Not-have disorders. Table 1 summarizes the main results by scale type and type of difference (with *P* levels based on 2-tailed *t* tests). Note that numbers refer to severity so that larger numbers are worse.

The results showed a clear Not-have–Have discrepancy, when we combined the results from all 3 scales (which was possible because all had been standardized). Have was considered less severe than Not-have. Across subjects, combining Self and Other, the mean *z*-score difference between Not-have and Have was 0.13 ($t_{74} = 2.23$, $P = 0.0291$, across subjects—these comparisons are based on 75 subjects, because a subject had to have at least 1 disorder of those we analyzed to contribute to the results). The effect was present for both Self (difference = 0.11, $t = 1.83$, $P = 0.0728$) and Other (difference = 0.14, $t = 2.27$, $P = 0.0260$). It is apparent that the Self-Other by Not-have–Have interaction did not appear. The effect was at least as large for Other as for Self.

Our main hypothesis was that the Not-have–Have discrepancy would be larger for the more vague scales. This was not supported. Table 1 shows the relevant results. Although the discrepancy was significant for the anchored and happiness scale and not for the vague scale, the 3 scales did not differ significantly in the size of the discrepancy. Importantly, the discrepancy was found for the anchored scale.

Self-Other Difference

To examine the Self-Other difference, we used all health disorders (since this difference was between questions presented on the same screen and thus had less error). First, we looked for an overall Self-Other difference across all 3 measures. To do this, we used the actual ratings. We did not need to standardize because the Self and Other responses were made to the same cases and could thus be compared directly. However, we multiplied the happiness ratings by -20 to roughly equate the scales (because most ratings were between 0

Table 2 Relationship between Scale Type and Self-Other Difference, in Severity Scores on a Rough 100-Point Scale for All Disorders, Experiment 1

Scale Type	Self	Other	<i>P</i>
Anchored	38.7	41.2	0.0028
Vague	55.9	56.8	<i>ns</i>
Happiness	44.9	47.6	0.0609
Overall	46.5	48.5	0.0105

Higher numbers represent worse health.

and -5). Table 2 shows the results for each scale (again, with high numbers representing worse health).

Combining the 3 measures, the overall difference amounted to a mean of 2.0 points on the 100-point scale, with ratings more severe for Other than for Self. As shown in Table 2, the difference was significant overall and for the anchored scale. Again, the differences among scales in the size of the effect were not significant.

It thus appears that the Self-Other discrepancy exists as hypothesized, and it therefore may account somewhat for the Have–Not-have discrepancy when Haves are asked about themselves and Not-haves are asked about others.

Although we had no particular hypothesis about the interaction between scale type and Other-Self, it appears, again, that the anchored scale is most sensitive to the difference.

Differences among Disorders

The Not-have–Have discrepancy depended on the disorder. To show this, we asked whether disorders with a higher discrepancy in one half of the subjects were also higher in the other half. This is in essence a test of split-half reliability of the measure of utility of disorders. We measured the mean discrepancy for the odd-numbered subjects for each disorder (combining all 3 measures) and the mean for the even-numbered subjects. The 2 sets of means were correlated across the 15 disorders ($r = 0.74$, $P = 0.0013$, 1 tailed). Table 3 shows, in the rightmost column, the Not-have–Have discrepancy for the different disorders. Although we made an effort to find all the usable common disorders, it seems that the overall result of a positive discrepancy could be an artifact of our sample—given the fact that disorders differed—and a different sample might have yielded even a reversed discrepancy. For example, migraines seem to be worse to those who have them than to those who do not. Figure 1 shows the Not-have–

Table 3 Mean Severity Ratings for the 11 Common Disorders and Self-Other (positive for Other worse than Self) and Not-have–Have Differences (positive for Not-have worse than Have), Experiment 1

Health Disorder	Number Having	Mean Severity ^a	Self-Other Difference ^a	Not-have–Have Difference ^a	Not-have–Have Difference z Score
Smoking	23	41.9	−14.0	15.0	0.41
Acne	13	43.6	4.2	14.0	0.39
Short stature	29	25.5	12.1	8.5	0.32
Insomnia	22	37.3	3.2	2.1	0.09
Asthma	11	49.2	3.2	2.7	0.09
Arthritis	18	56.5	0.8	1.2	0.05
Nearsightedness	43	21.7	6.5	1.6	0.04
Back problems	21	65.1	−1.7	−0.8	−0.01
Overweight	10	56.0	0.5	−0.8	−0.02
Knee	12	38.1	4.0	−3.2	−0.11
Migraines	22	56.4	−0.8	−3.9	−0.15

Ordered by discrepancy.
a. Rated on a 0 to 100 scale.

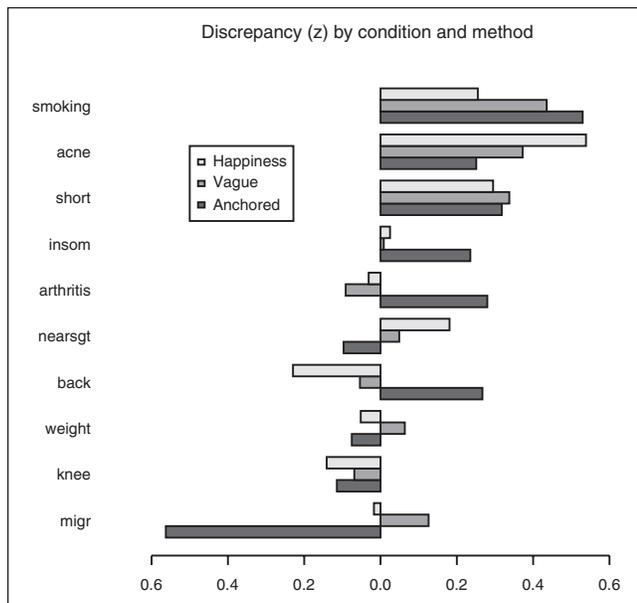


Figure 1 Have–Not-have discrepancy by disorder and scale type, experiment 1. Positive numbers indicate that Not-haves rate the disorder as worse than Haves do. (Note that migraines are rare.)

Have differences of standardized scores as a function of disorder and scale type.

We found the same kind of consistency across disorders for the Self-Other difference ($r = 0.89$). The means for the disorders are shown in Table 3. Table 3 also shows the mean severity ratings (averaging all 3 measures, with happiness multiplied by -20 before averaging).

EXPERIMENT 2

The main result of experiment 1 was a failure to find a larger discrepancy between Haves and Not-haves in the vague scale or in the happiness scale than in the anchored scale. The happiness scale did, however, show a slightly larger discrepancy than the other 2 scales did.

It is possible that a scale focusing more explicitly on QOL, rather than severity of a health state, might show a larger discrepancy: Especially when the reference points are vague, people with a disorder may tend to think about others with the same disorder when they evaluate their QOL.

Experiment 2 (actually done after experiment 3) involved 4 types of screens for each condition in a 2×2 design: severity (of a disorder) versus WOL crossed with vague versus specific. Each screen had a Self question and an Other question. And, of course, we compared Not-have and Have conditions.

Method

Ninety-nine subjects completed a questionnaire on the World Wide Web. Their ages ranged from 19 to 68 years (median = 36); 22.2% were male, and 11% were students. (Twenty-three percent of these subjects had previously done experiment 1 or experiment 3 or both; 38% of those who did experiment 3 had previously done experiment 1.) The questionnaire began as follows:

Health Conditions

The following study concerns judgments of chronic health conditions. There are 40 questions (screens). Each question presents a short description of a health condition and asks one of two types of rating question.

One question concerns health. We ask you to rate the health condition on a 50-100 scale. Questions differ in how 50 and 100 are defined. Pay attention to these definitions.

The Other question concerns the overall quality of life of those who have the condition.

Note: You can go below 50 if you feel that a condition is worse than the definition of "50."

You will make the ratings from two points of view. One is your own point of view, supposing that you have the condition. If you have really had the condition, please pay attention to the description of the condition, which may be more or less severe than your own case.

The other point of view is that of the average person who gets the condition described.

In each case, imagine that either you or the average person has had the condition for at least 6 months. Also, the condition will not change in the foreseeable future. It will not get better, and it will not get worse. Do not suppose that it will improve.

A typical question, in the QOL-specific condition, appeared as follows (with the definitions of 100 and 50 in green and red, respectively):

Suppose that *you* had the condition: **acne—pimples all over face**.

For yourself, on a scale of overall **quality of life** in which 100 is as good as that of someone with a meaningful job, friends, family, and good health, and

50 is as bad as that of someone who cannot walk more than 10 feet because of partial paralysis, has a dull job, and no close family or friends,

where would you put **acne**?

Now suppose that *the average person* had **acne (pimples all over face)**.

Where do you think that the average person would put **acne** on the same scale?

The scale definitions for the QOL-vague condition were as follows: "100 is a very good quality of life and 50 is a very poor quality of life." (We used 50 to make it easier for subjects to assign a number below the bottom anchor.) For the health-specific condition they were "100 is as healthy as a 20-year-old with no health conditions and 50 is as severe as being unable to walk more than 10 feet because of partial paralysis," and for the

health-vague condition they were "100 is very good health and 50 is very poor health." The 4 conditions were blocked and presented in a different random order for each subject.

The health states were chosen on the basis of previous studies to be common. Their order was randomized separately for each subject. Their definitions were as follows:

Asthma	attacks of breathing difficulty—1 per week
Back pain	pain from lifting objects
Insomnia	2 hours less sleep than desired on most nights
Shortness	6 inches shorter than average height for sex
Overweight	inability to lose excess weight
Nearsightedness	glasses required
Acne	Pimples all over face
Smoking habit	Uncontrollable addiction to cigarettes
Arthritis	Pain in hips or shoulders with some movements
Heart disease	Occasional chest pain from climbing stairs

At the end of the study, subjects again indicated which disorders they had for at least a year.

Results

Table 4 shows the mean ratings for the 4 conditions (after elimination of bad data, as we shall describe). Here, high numbers represent good health or good QOL. It is apparent that the 4 measures agreed closely on the relative severity of the health states, and the health states vary considerably in severity. Some subjects seemed to misunderstand the scale for some condition blocks. To assess misunderstanding, we correlated each subject's scores in each block with the 10 means based on all 4 conditions. We eliminated blocks when the correlation was less than 0.25. This resulted in deletion of 16.3% of the data. Two subjects had all their data deleted, and 58 had no data deleted. All statistical analyses are based on whatever data were not eliminated.

As is apparent in Table 4, Self ratings are consistently higher (less severe) than Other ratings, as found in experiment 1 (mean difference of 2.01, $t_{57} = 6.77$, $P = 0.0000$). Specific ratings are also higher than vague ratings (mean difference = 3.02, $t_{57} = 5.68$, $P = 0.0000$), but this is surely the result of the specific scale definitions, not a general result. Health and QOL ratings did not dif-

Table 4 Mean Ratings in Experiment 2 for Self and Other

	Health		Quality of Life	
	Specific	Vague	Specific	Vague
Nearsighted				
Self	92.1	92.3	91.0	90.1
Other	89.5	87.9	88.8	88.0
Short stature				
Self	90.7	90.0	90.9	90.1
Other	86.1	85.3	86.6	85.5
Acne				
Self	87.1	83.4	88.3	84.2
Other	82.3	78.4	82.9	80.8
Insomnia				
Self	83.5	79.4	82.3	79.8
Other	80.7	77.1	81.1	78.9
Back problems				
Self	81.8	75.6	77.9	76.4
Other	78.8	74.6	75.9	75.5
Asthma				
Self	80.2	76.5	76.9	73.8
Other	77.2	73.2	74.8	72.5
Overweight				
Self	78.3	77.3	76.8	72.3
Other	75.2	73.5	74.9	70.0
Smoking				
Self	80.4	72.4	74.3	70.4
Other	78.3	73.6	75.9	71.6
Arthritis				
Self	77.8	74.4	73.8	72.5
Other	74.8	71.4	72.2	71.6
Heart disease				
Self	74.7	71.3	70.6	67.8
Other	71.4	68.6	68.6	66.2
Mean				
Self	82.7	79.3	80.3	77.8
Other	79.4	76.4	78.2	76.1

Based on cleaned-up data. Higher numbers represent better health.

fer significantly. Self-Other difference was slightly greater for health than for QOL ($t_{57} = 2.55, P = 0.0135$); this too does not seem to imply any general conclusion.

Figure 2 shows the results for the Not-have versus Have discrepancy for the 10 health states and the 4 conditions (for Self and Other combined); positive numbers indicate that Not-haves rate the disorder as worse than Haves do. We did not find an overall Not-have-Have discrepancy, even for Self. This was because the health states differed in the direction of this effect, as is apparent from Figure 2. The correlation across disor-

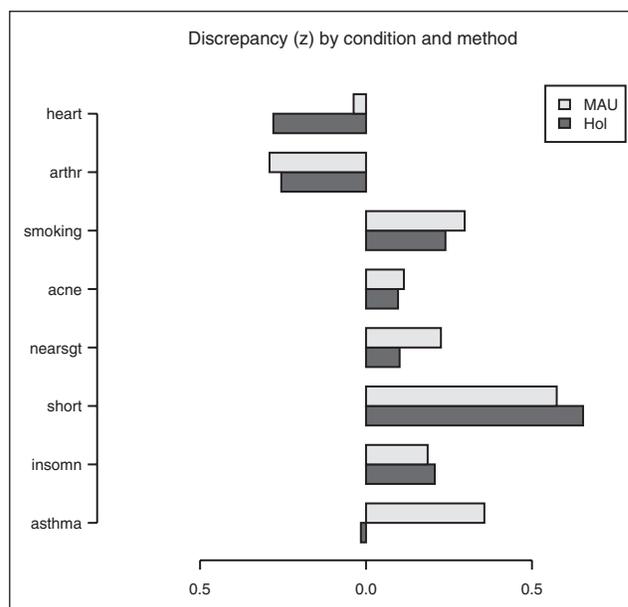


Figure 2 Have-Not-have discrepancy by disorder and method, experiment 2. Positive numbers indicate that Not-haves rate the disorder as worse than Haves do. V = vague; S = specific; Q = quality of life; H = happiness.

ders between the discrepancy measures for 2 halves of the sample (computed as in experiment 1) was 0.49 ($P = 0.0771, 1$ tailed).

Although the overall discrepancy (Have less severe than Not-have) was not significant, we note, first, that it was greater for Self (0.063) than for Other (0.030, in contrast to experiment 1, $t_{54} = 2.22, P = 0.0303$).

It is also apparent (from Figure 2) that the discrepancy, to the extent to which we find it, was no larger for QOL than for health. In fact, it was nearly significant for health alone in the Self condition (mean = 0.25, $t_{69} = 1.92, P = 0.0596$) and in the opposite direction for QOL (-0.17), resulting in a significantly greater discrepancy (Have higher than Not-have) for health than for QOL ($t_{63} = 3.03, P = 0.0036$). This result starkly contradicts the hypothesis that the discrepancy would be greater for QOL.

Similarly, the discrepancy was no larger for vague than specific. Again, we found the opposite ($t_{63} = -2.23, P = 0.0291$): The Not-have-Have discrepancy was greater (higher ratings for Have) in the specific than in the vague condition.

EXPERIMENT 3

The purpose of experiment 3 was to ask whether the discrepancy effect was the result of a focusing illusion. The idea of a focusing illusion is that when people are

asked about a difference between 2 disorders, they focus on the attributes of those disorders that are different, because of the way the question is asked. They thus exaggerate the magnitude of the difference. For example, people asked to compare life in California with life in the Midwestern states of the United States focused on the weather, thus overestimating the benefits of living in California.¹⁵ Likewise, people who do not have a disorder could focus on the attributes of life that are more affected by that disorder.

As we explained in the introduction, a previous study¹⁶ found no evidence for a focusing illusion as an explanation of the Have–Not-have discrepancy. The study tried to reduce such an illusion, if it existed, by calling subjects' attention to a variety of attributes that characterize the goodness of life. Asking subjects how a disorder affected each of these dimensions did not affect their subsequent rating of the same disorder.

Ubel and others¹⁶ used rare conditions, such as below-the-knee amputation and paraplegia, so they were unable to make the Not-have–Have comparison in their sample (jurors). We do not know whether they would have found a discrepancy effect if they had done so. Also, it is possible that subjects may have understood that the disorders did not affect all life attributes equally yet still reverted to the focusing illusion even after they were forced to consider all the attributes.

In the present experiment, we followed the basic design of Ubel and others.¹⁶ In particular, we first asked for holistic ratings, then we asked for attribute-by-attribute ratings, and finally we asked for holistic ratings again. However, we asked about the attribute-by-attribute ratings in a way that allows us to do a rough calculation of utility based on multiattribute utility (MAU) theory.²¹ We thus call this an MAU elicitation. In particular, we asked for numerical ratings on each of 7 attributes, each with a clearly anchored endpoint. At the end of the experiment, we asked for weights of these attributes. To calculate the MAU of each disorder for each subject, we multiplied each attribute rating by the weight of that attribute and then added up these products across the 7 attributes.

If the Not-have–Have discrepancy results from a focusing illusion, it should largely disappear in the MAU ratings because subjects were forced to rate attributes one at a time. Thus, even if the effect of this forced variety is limited to the MAU task itself, we should be able to detect it. We do not need to rely on transfer to the subsequent holistic rating task.

For the holistic question, we used paralysis from the waist down as the standard, rather than death, because we thought that people might be reluctant to say that a disorder they had was anywhere near as bad as death.

Method

Seventy-nine subjects completed a questionnaire on the World Wide Web. Their ages ranged from 18 to 74 years (median = 36); 33% were male, and 14% were students.

The questionnaire began as follows:

Preference for Health Conditions

This study concerns judgments of chronic health conditions. It has 24 screens. Please read all of these instructions carefully.

In each question, you will see a short description of a health condition, and you will answer one of two types of questions.

One question concerns undesirability, which means the strength of preference for not having the condition. This type of question will come first, and then it will be repeated at the end. Do not worry about whether your answer is the same or not. Just try to answer accurately both times.

You answer this question on a scale where 0 means “not having the condition” and 100 means “as bad as being paralyzed from the waist down.” You can use numbers greater than 100 if necessary (but no greater than 200).

The other type of question concerns the negative effect of the condition on several domains of your life:

1. Pain and discomfort
2. Economic standard of living
3. Work
4. Love life
5. Family life other than love life
6. Spiritual life broadly defined
7. Leisure activities other than family life

Please try to interpret these descriptions so that they do not count the same effects twice. For example, if “spiritual life” includes communing with nature, do not also count this as part of “leisure.”

You answer these questions on a scale where 0 means “no negative effect” and the meaning of 100 is specified in the question. If you think that some health condition has a **positive** effect, then use a **negative number** for your response. Remember, this question is about negative effects.

At the end, you will be asked a few additional questions.

Each question asks you to “suppose that you had” the condition. Of course, you may actually have it now. If so, please pay attention to the description, which

may be more or less severe than your own case. Rate that description, not your own case.

In each case, imagine that you have had the condition for 6 months. Also, the condition will not change in the foreseeable future. It will not get better, and it will not get worse. This is important. Do not suppose that it will improve.

We used the following disorders, which had showed a discrepancy in the hypothesized direction in experiment 1 (except for heart disease, which was modified to be less severe).

Asthma	Attacks of breathing difficulty, 1 per week
Insomnia	2 h less sleep than desired on most nights
Short stature	6 inches shorter than average for gender
Nearsightedness	Glasses required
Acne	Pimples all over face
Smoking habit	Pack a day of cigarettes
Arthritis	Pain in hips or shoulders with any movement
Heart disease	Occasional chest pain from climbing stairs

For the holistic task, a typical question read as follows:

Suppose that you had the condition: arthritis—pain in hips or shoulders with any movement. On a scale of overall undesirability in which 0 is not having arthritis (with everything else the same) and 100 is waist-down paralysis, where would you put arthritis?

For the MAU task, a typical question read as follows:

Suppose that you had the condition: insomnia—2 hours less sleep than desired on most nights. Rate this condition for its negative effect on each of the following domains of your life. 0 means “no negative effect at all.” 100 is defined for each question. (Use a negative number for a positive effect.)

1. Pain, fatigue, and discomfort: 0 = no effect; 100 = as bad as death.
2. Economic standard of living: 0 = no effect; 100 = dire poverty.
3. Work: 0 = no effect; 100 = unable to do any work.
4. Love life: 0 = no effect; 100 = love life nonexistent.
5. Family life: 0 = no effect; 100 = family life nonexistent.
6. Spiritual life: 0 = no effect; 100 = spiritual life nonexistent.

7. Leisure activities: 0 = no effect; 100 = activities nonexistent.

The order of health conditions was randomized separately for each subject and then fixed for the three parts of the questionnaire.

The weight elicitation, at the end of the questionnaire, read as follows:

Now please rate each of the following on a scale where 0 represents “not bad at all” and 100 represents “as bad as paralysis from the waist down.” When you make these ratings, think only about the thing you are rating. Try to imagine that everything else is the same. (This is hard, but do your best.)

Then subjects saw a list of all the lower ends of each scale in the MAU task, for example, “pain, fatigue, and discomfort as bad as death.”

Results

Seven subjects were dropped because they gave 0 responses to all or most of the holistic questions in one section (including the more severe disorders) or because they gave 0 to all the questions about weights. Also, some subjects were dropped from some analyses. In particular, subjects whose holistic ratings before and after the MAU ratings did not correlate positively with each other across disorders were dropped for analyses of the holistic ratings. When holistic ratings and MAU ratings correlated negatively, we dropped the one with the lowest correlation with the mean of all severity measures across disorders. By these criteria, we dropped 1 subject from analysis of the MAU data and 13 subjects from analysis of the holistic data. The first holistic task seemed particularly prone to induce reversed judgments (high numbers for less severe disorders).

To compute MAU utilities, we first rescaled the weights for each subject so that the maximum weight was 1. Likewise, we rescaled the ratings so that the highest rating given by each subject was 1. Then we multiplied the weights by the ratings for each attribute for each disorder for each subject. Notice that the weights were elicited using the same endpoints as those used in the rating task.

To assess the Not-have–Have discrepancy, we used the standardized responses as in experiment 1. We also standardized the calculated MAU utilities by disease, to ensure that a Not-have–Have discrepancy would not result artifactually from a negative correlation between severity and frequency of conditions.

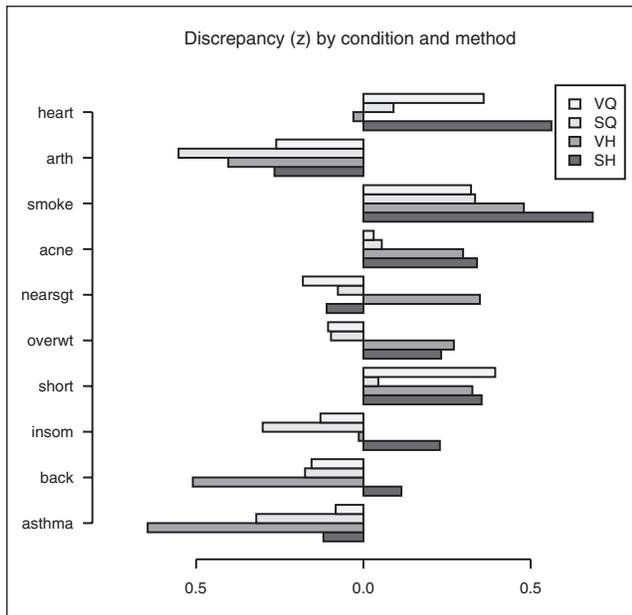


Figure 3 Have–Not-have discrepancy by disorder and method, experiment 3. Positive numbers indicate that Not-haves rate the disorder as worse than Haves do.

The Not-have–Have discrepancy was significant overall, averaging the holistic (averaged over the 2 parts) and MAU ratings (mean difference = 0.16 in terms of standardized scores, $t_{61} = 2.19$, $P = 0.0320$, 2 tailed; positive numbers mean that Have is less severe than Not-have). It was also significant for the MAU ratings (mean = 0.15, $t_{72} = 2.51$, $P = 0.0144$) but not for the holistic ratings (mean = 0.14, $t_{64} = 1.57$, $P = 0.1213$; the fact that the overall mean was higher is the result of omission of different subjects for different comparisons and rounding). These results are a clear rejection of the focusing hypothesis, which predicted an effect for the holistic ratings but not for the MAU ratings.

Again, the discrepancy varied across disorders, as shown in Figure 3. The split-half correlation across the 8 disorders in the discrepancy (computed as in experiments 1 and 2) was 0.70 ($P = 0.0257$, 1 tailed, with $df = 6$).

The pattern of attribute ratings should vary by disorder. To test whether this was true, we performed an analysis of variance of the attribute ratings, using subject, attribute, and disorder as factors. Importantly, the interaction between attribute and disorder was significant ($F_{42,3276} = 28.05$, $P = 0.0000$). Main effects of attribute and disorder were also large and significant. Table 5 shows the mean ratings, on a scale in which the worse

Table 5 Mean Ratings and Discrepancies on the Same Scale as a Function of Attribute and Disorder, Experiment 3

	Pain	Economics	Work	Love	Family	Spirit	Leisure
Mean disutility ratings							
Asthma	0.49	0.20	0.31	0.22	0.18	0.09	0.42
Insomnia	0.39	0.19	0.41	0.27	0.30	0.14	0.37
Short stature	0.10	0.13	0.13	0.32	0.12	0.07	0.15
Nearsightedness	0.10	0.09	0.09	0.08	0.08	0.07	0.15
Acne	0.22	0.15	0.21	0.59	0.19	0.09	0.24
Smoking	0.33	0.37	0.21	0.31	0.28	0.16	0.29
Arthritis	0.74	0.41	0.60	0.44	0.31	0.11	0.62
Heart disease	0.64	0.29	0.48	0.38	0.30	0.12	0.52
Mean Not-have–Have discrepancy in disutility							
Asthma	0.08	−0.02	0.06	−0.03	−0.01	−0.07	0.01
Insomnia	−0.05	0.02	0.02	0.17	−0.04	−0.08	0.01
Short stature	0.03	0.09	−0.07	0.00	0.11	0.05	0.02
Nearsightedness	−0.02	−0.04	0.06	0.00	0.00	0.06	−0.04
Acne	−0.01	0.02	0.02	−0.02	0.01	−0.08	0.03
Smoking	−0.08	−0.01	0.03	0.10	0.10	−0.01	−0.02
Arthritis	0.08	−0.02	−0.03	−0.08	−0.02	0.06	−0.14
Heart disease	0.06	0.11	0.05	0.03	−0.05	−0.01	−0.12

Worst = 1.0; no effect = 0. Higher numbers represent worse health.

end of each attribute is 1 and “no effect” is 0, and the Not-have–Have discrepancy on the same scale. The interaction between attribute and disorder is illustrated by the high numbers for the effect of arthritis on pain and the effect of acne (and shortness) on love life.

We examined the Not-have–Have discrepancy for each attribute. The largest discrepancies were those for family life ($t_{72} = 3.18, P = 0.0022$, 2 tailed), love life ($t = 2.10, P = 0.0396$), and spiritual life ($t = 3.53, P = 0.0007$). The effect for work was nearly significant ($t = 1.75$). No other effect was significant by the usual criterion, although all were in the hypothesized direction, including pain and discomfort. These results suggest that the Not-haves may fail to recognize how little effect the disorders have on some domains.

DISCUSSION

We have provided evidence against 2 plausible accounts of the Have–Not-have discrepancy. By one account, the discrepancy effect is the result of shifting reference points. If this is true, the effect should be reduced by defining the reference points more clearly. If anything, such definition increased the effect. By another account, the discrepancy is due to a focusing illusion, in which those who do not have a disorder focus on its effects and ignore what it does not affect. We provided evidence against this account by showing that the discrepancy was not reduced by asking subjects to rate effects on several domains, some of which were not affected by the disorders in question.

Our results also indicate that the discrepancy can be studied with Web respondents with common disorders. Although each disorder was present in a minority of subjects, most subjects had at least one disorder. For more general purposes, the list could be expanded to include other aspects of life aside from health, such as being single, poor, or unemployed or good things such as being wealthy. Indeed, we may have done this by including shortness as a disorder. We have no reason to think that the general principles underlying the discrepancy would differ for nonhealth states.

We found somewhat consistent differences between disorders in the magnitude and direction of the discrepancy (although this finding was not quite significant in experiment 2). One possible explanation of these differences—and there may well be more than one—is that the usual discrepancy (Not-have worse than Have) is found for disorders that have an external manifestation, visible to others, whereas the opposite is found for disorders that involve pain or unexpressed emotion. Further research should examine these differences among disorders.

Our main conclusion is that vague scales are not the cause of the discrepancy. If anything, the discrepancy is larger when scales have more precise anchors. Thus, the discrepancy is not a simple artifact of the use of vague scales. (Note, however, that this negative result does not justify the use of vague scales for other purposes, such as comparison of attribute weights.)

A limitation of this conclusion stems from our use of within-subject designs. Those who have a disorder might recalibrate their scale for judgments of all disorders (or of their absence). If, for example, a paraplegic rated himself as “happy,” he might still believe that others are happier and rate a typical person without paraplegia as “ecstatic.” We do not think that such a general recalibration is likely for the kinds of scales we used, for the kinds of minor disorders that we studied, but such recalibration may exist elsewhere. If it exists, a different design is required to detect it.

We also found further evidence that defocusing manipulations do not reduce the discrepancy. Unlike earlier studies, these results did not rely on the carryover from a defocusing manipulation to another task. We found the discrepancy in the MAU task itself.

The focusing hypothesis is not quite dead, however. It is possible that, even within a life domain, people with a disorder focus on subattributes of that domain that are less affected by the disorder. It may be difficult to draw a line between “attributes” and “activities,” however. It may therefore be difficult to distinguish this form of the focusing hypothesis from other hypotheses that involve knowledge of specific adaptations (such as using a computer for reading when one is blind).

Finally, people think they will adapt better than others. In experiment 2, this is more true when they have in fact adapted (Haves). This result could explain some of the discrepancy found in previous studies, if people who do not have the disorder think about others rather than themselves. For example, Jansen and others found that experience with chemotherapy did not change utilities assigned to a chemotherapy scenario, but patients having chemotherapy rated the scenario as less severe than did other patients before, during, and after the chemotherapy itself.²² Perhaps the nonchemotherapy patients were rating the scenario for others. The Self-Other difference could also explain the finding that impersonal scenarios are rated as more severe than personal ones.²³

The main remaining explanations of the discrepancy are failure of Not-haves to predict adaptation,²⁴ Haves’ self-deception, and Haves’ loss aversion. The differences among these explanations matter. Adaptation is real, and self-deception is, in an important

sense, false. We may think of people's values in terms of what Keeney calls fundamental values and means values.²⁵ Fundamental values are what ultimately matter to each person. Means values are what matter because people believe they are means to satisfying fundamental values. If beliefs are false, then the means values lose their claim to represent a person's good.²⁶ Thus, when we help people make decisions, we do not do them any good when we honor their values based on self-deception and when these conflict with their fundamental values. If people with disabilities are deceiving themselves about how close to normal they are, then, to this extent, we should discount their judgments and work harder to cure and prevent their disorders than their own judgments would imply.

In the case of loss aversion, Not-haves could see the acquisition of a disorder as a loss, whereas Haves see the end of the disorder as a gain, and gains are weighed less heavily than losses. In this case, the Not-haves would again have the more accurate view, if utility judgments were compared to those of other losses (of health or money).

On the other side, misprediction of adaptation—possibly a result of anchoring on the initial change and underadjusting—is a false belief held by Not-haves. To the extent to which adaptation is real and not predicted, we should discount these judgments in evaluating the severity of disorders.

Of course, all 3 kinds of error could happen. And our results suggest a different kind of Not-have error as well, a failure to appreciate the severity of disorders that have no external manifestation.

REFERENCES

1. Brickman P, Coates D, Janoff-Bulman R. Lottery winners and accident victims: is happiness relative? *J Pers Soc Psychol.* 1978;36:917–27.
2. Brickman P, Campbell DT. Hedonic relativism and planning the good society. In: Appley MH editor. *Adaptation-level theory: a symposium.* New York: Academic Press; 1971.
3. Boyd NF, Sutherland HJ, Heasman KZ, Trichter DL, Cummings BJ. Whose utilities for decision analysis? *Med Decis Making.* 1990;10:58–67.
4. Sackett DL, Torrance GW. The utility of different health states as perceived by the general public. *J Chron Dis.* 1978;31:697–704.
5. DeWit GA, Busschbach JVV, DeCharro FTH. Sensitivity and perspective in the evaluation of health status: whose values count? *Health Econ.* 2000;9:109–96.
6. Llewellyn-Thomas HA, Sutherland HJ, Theil EC. Do patients' evaluations of a future health state change when they actually enter that state? *Med Care.* 1993;31:1002–12.
7. O'Connor AMC, Boyd NF, Warde P, Stolbach L, Till JE. Eliciting preferences for alternative drug therapies in oncology: influence of treatment outcome description, elicitation technique and treatment experience on preferences. *J Chron Dis.* 1987;40:811–8.
8. Sprangers MAG, Schwartz CE. Integrating response shift into health-related quality-of-life research: a theoretical model. *Soc Sci Med.* 1999;48:1507–15.
9. Visser MR, Smets EM, Sprangers MA, de Haes HJ. How response shift may affect the measurement of change in fatigue. *J Pain Symptom Manage.* 2000;20:12–8.
10. Bernhard J, Hürny C, Maibach R, Herrmann R, Laffer U. Quality of life as subjective experience: reframing of perception in patients with colon cancer undergoing radical resection with or without adjuvant chemotherapy. *Ann Oncol.* 1999;10:775–82.
11. Bernhard J, Lowi A, Maibach R, Hürny C. Response shift in the perception of health for utility evaluation: an explorative investigation. *Eur J Cancer.* 2001;37:1729–35.
12. Adang EMM, Kootstra G, Engel GL, van Hooff JP, Merckelbach HLGJ. Do retrospective and prospective quality of life assessments differ for pancreas-kidney transplant recipients? *Transplant Int.* 1998;11:11–15.
13. Postular D, Adang EM. Response shift and adaptation in chronically ill patients. *Med Decis Making.* 2000;20:186–93.
14. Jansen SJ, Stiggelbout AM, Wakker PP, Nooij MA, Noordijk EM, Kievit J. Unstable preferences: a shift in valuation or an effect of the elicitation procedure? *Med Decis Making.* 2000;20:62–71.
15. Schkade DA, Kahneman D. Does living in California make people happy? A focusing illusion in judgments of life satisfaction. *Psychol Sci.* 1998;9:340–6.
16. Ubel PA, Loewenstein G, Hershey JC, et al. Do nonpatients underestimate the quality of life of chronic health states because of a focusing illusion? *Med Decis Making.* 2001;21:190–9.
17. Ubel PA, Loewenstein G, Jepson C. Quality of life ratings of unfamiliar disabilities: the contribution of a focusing illusion and failure to predict adaptation. Unpublished manuscript, 2002.
18. Weinstein ND. Unrealistic optimism about future life events. *J Pers Soc Psychol.* 1980;39:806–20.
19. Andresen EM, Vahle VJ, Lollar D. Proxy reliability: health-related quality of life (HRQoL) measures for people with disability. *Qual Life Res.* 2001;10:609–19.
20. Todorov A, Kirchner C. Bias in proxies' reports of disability: data from the National Health Interview Survey on Disability. *Am J Public Health.* 2000;90:1248–53.
21. Keeney RL, Raiffa H. *Decisions with multiple objectives: preference and value tradeoffs.* New York: Cambridge University Press; 1993.
22. Jansen SJT, Kievit J, Nooij MA, Stiggelbout AM. Stability of patients' preferences for chemotherapy. *Med Decis Making.* 2001;21:295–306.
23. Chapman GB, Elstein AS, Kuzel TM, et al. Prostate cancer patients' utilities for health states: how it looks depends on where you stand. *Med Decis Making.* 1998;18:278–86.
24. Gilbert DT, Pinel EC, Wilson TD, Blumberg SJ, Wheatley TP. Immune neglect: a source of durability bias in affective forecasting. *J Pers Soc Psychol.* 1998;75:617–38.
25. Keeney RL. *Value-focused thinking: a path to creative decisionmaking.* Cambridge, MA: Harvard University Press; 1992.
26. Baron J. Norm-endorsement utilitarianism and the nature of utility. *Econ Philosophy.* 1996;12:165–82.