

The Metrics of Subjective Wellbeing Data: An Empirical Evaluation of the Ordinal and Cardinal Comparability of Life Satisfaction Scores

Ingebjørg Kristoffersen¹ 

Accepted: 1 December 2015 / Published online: 15 December 2015
© Springer Science+Business Media Dordrecht 2015

Abstract This paper is motivated by the lack of consensus on the metrics of subjective wellbeing measurement scales. Subjective wellbeing data are frequently treated as though they are cardinally comparable both across and within individuals, though very little evidence exists to support these assumptions. Because wellbeing cannot be observed directly, cardinality must remain an assumption, which is usually imposed based on statistical convenience rather than on reason. The premise of this paper is that it is both possible and useful to make this assumption more informed. The analysis applies the principle of simultaneous conjoint measurement to improve our understanding of what information is contained within subjective wellbeing scores. Specifically, the metrics of the eleven-point numeric life satisfaction scale is evaluated using the MH5 mental health survey instrument. Under the assumption that the response function for MH5 is identifiable by the Rasch model, the shape of the response function for life satisfaction is potentially observable indirectly via the association between life satisfaction and MH5. The results presented here suggest life satisfaction scores are ordinally distinct, in terms of these mental health data, which supports the assumption of ordinal comparability. Under the aforementioned assumption, these scores are also approximately equidistant, which supports cardinal comparability. This pattern is found both across individuals and within individuals across time.

Keywords Life satisfaction · Subjective wellbeing · Mental health · Cardinality · Ordinality · Response functions · Methodology

✉ Ingebjørg Kristoffersen
inga.kristoffersen@uwa.edu.au

¹ School of Business, The University of Western Australia, Perth, Australia

1 Introduction

The use of survey-based measures of happiness and satisfaction, defined more broadly as subjective wellbeing, is common across the social sciences and has more recently become mainstream also in economics. However, scholars differ in their assumptions about the metrics of subjective wellbeing data. Subjective wellbeing data are typically gathered by asking people a single question about how happy or satisfied they are with their lives in general, or with aspects of their lives. These data are commonly treated as ordinally comparable, and increasingly also cardinally comparable, both across individuals and within individuals. Cardinal comparability requires a unique and linear relationship between true wellbeing or utility, and measured subjective wellbeing, though we have little information on which to base such an assumption. In a recent survey Hirschauer et al. (2014) identify this as a key issue which needs to be addressed in order to improve our approach to and use of subjective wellbeing data in economic analysis and policy applications. The purpose of this study is to address this issue.

The assumption of cardinal comparability is often justified largely on the basis of statistical requirements, and occasionally requirements for interpretation, rather than on reason. In particular, cardinality is often considered a necessary assumption due to the importance of individual fixed effects in explaining cross-sectional differences in subjective wellbeing data and their ability to absorb information which otherwise might cause bias (Ferrer-i-Carbonell and Frijters 2004). However, continued improvements in data availability and technology may ultimately challenge this type of justification.¹ The cardinality assumption ought in any case to be considered based on its reasonability and implications rather than just on its usefulness. So far, the most salient such justification appears to be that estimates of models for ordered discrete and continuous data tend to be highly consistent.² The premise of this study is that we both can and should do better.

There are two main approaches to interpreting subjective wellbeing data. A fully operationalised definition implies that subjective wellbeing is defined by the way it is measured. This might seem a convenient means of circumventing difficult questions of metrics, though it does not avoid problems of arbitrary measurement and bias due to scale restrictions. More commonly, subjective wellbeing data are treated as measures of true psychological wellbeing, and often specifically of utility in economic analyses. This approach therefore implies certain assumptions about the response function by which true wellbeing or utility translates into observed survey responses.

Since true wellbeing and utility are abstract (latent) psychological concepts which we cannot observe directly, we cannot observe the response function for subjective wellbeing. Oswald (2008) argues that subjective wellbeing data therefore cannot be treated as cardinal measures of utility, that we cannot use these data to estimate marginal utilities, and that we must restrict ourselves to estimating marginal rates of substitution, which is sufficient in many applications. Conversely, Layard et al. (2008) argue that the cardinality assumption is both reasonable and justified, and that accurate information about the curvature of the

¹ Accounting for individual fixed effects is important to the extent that we cannot yet capture the necessary missing information by other means, and cardinal comparability is important to the extent that this is—currently—a requirement for fixed-effects panel model estimation. Individual fixed effects have been shown to be accounted for, to a large extent, by personality traits, which are now captured in some survey data (Boyce 2010). Thus, our improved ability to capture this information may therefore potentially render fixed-effects panel model estimation less important.

² See for example Blanchflower and Oswald (2004, 2005), Gardner and Oswald (2001), Headey and Wooden (2004), Van Praag and Ferrer-i-Carbonell (2004), and Ferrer-i-Carbonell and Frijters (2004).

utility function is necessary for correct policy design. On closer consideration, the estimation of marginal rates of substitution also implies certain assumptions about the comparability of subjective wellbeing scores, for which there is currently little justifiable basis.

This paper evaluates the case for ordinal and cardinal comparability of the eleven-point life satisfaction scale by comparing data on life satisfaction and mental health from the Household, Income and Labour Dynamics in Australia (HILDA) survey. The observed association between life satisfaction and mental health may be interpreted in different ways, depending on what assumptions are considered reasonable. A key feature of this analysis is that this association is treated as a combination of two separate response functions, which each taps into the same latent concept of wellbeing or utility, in the spirit of simultaneous conjoint measurement (Luce and Tukey 1964).

The rationale behind this particular approach is that, if we know something about the response function for mental health, the association between life satisfaction and mental health determines the outer boundaries of what form the response function for life satisfaction might take. If the response function for mental health can be assumed to be linear, as intended by the way in which this measure is constructed, then the shape of the response function for life satisfaction is potentially directly observable via the association between life satisfaction and mental health. These assumptions are considered carefully in turn.

The paper proceeds as follows. Section 2 provides an a priori evaluation of the basic assumptions underpinning common uses of subjective wellbeing data. This sets the scene for the empirical analysis that follows. Section 3 describes the method of analysis, and Sect. 4 presents the data and the results. Finally, Sect. 5 provides a summary and discussion of the findings and their potential implications.

2 Subjective Wellbeing Data and the Problem of Metrics

The instruments used to measure subjective wellbeing place obvious constraints on the types of assumption that can reasonably be made about these data. Early surveys have tended to ask respondents to report their happiness or satisfaction by selecting one of a set of ordered responses. For example, the US General Social Survey asks “Taken all together, how would you say things are these days: would you say you are very happy, pretty happy, or not too happy?” The Eurobarometer survey asks “On the whole, are you very satisfied, fairly satisfied, not very satisfied, or not at all satisfied with the life you lead?” It is difficult to justify an assumption that these responses contain information beyond order.

The use of numeric scales, rather than ordered responses, is now used in many large national surveys, including the HILDA Survey, the German Socio-Economic Panel (GeSoEP) survey and the World Values Survey (WVS). Respondents are then asked to indicate their happiness or satisfaction by selecting an integer on a numeric scale, with verbal anchors at both ends and often also in the middle (to indicate neutrality). This approach appears to convey some intention of cardinality.

2.1 Basic Assumptions and Implications for Data Use

The types of assumptions made about subjective wellbeing data in economic and other analyses can be described as follows:

1. First and foremost, some degree of *construct validity* is required: a subjective wellbeing measure must succeed in capturing the necessary information about the

- relevant underlying (latent) variable of interest, such as psychological wellbeing or utility. In practical terms, we must have reason to assume a positive (monotonic) association between subjective wellbeing and true wellbeing or utility.³
2. Ordinal comparability of subjective wellbeing requires that the measurement scale is *unique* in terms of true wellbeing or utility. That is, a given score or response must infer the same (or sufficiently similar) wellbeing or utility across individuals or within individuals, or both. This is a basic requirement for most uses of subjective wellbeing data.
 3. Cardinal comparability further requires *equidistance* of score-points or ordered responses.⁴ That is, the differences between adjacent score-points or responses are interpreted as equal (or sufficiently similar) in terms of true wellbeing or utility. This, in turn, requires that wellbeing or utility is itself cardinal, and furthermore that utility or wellbeing is bounded in the same way (or approximately the same way) as is the measurement scale.

Ordinal comparability is a basic requirement for most uses of subjective wellbeing data, also when used to estimate marginal rates of substitution. If we are only interested in observing changes in subjective wellbeing scores within individuals across time (or across categories), then the condition of uniqueness is not strictly necessary, though the condition of equidistance then often is. If we are interested in comparing scores across individuals, as is most often the case, then the condition of uniqueness (or approximate uniqueness) is a minimum requirement, and cardinal comparability is often also implied.⁵ Cardinal comparability, specifically with respect to utility, is necessary for the estimation of marginal utility in the manner of Layard et al. (2008).

2.2 Fundamental Measurement Issues

2.2.1 Construct Validity

A large literature has emerged over the last few decades to demonstrate that people who report greater subjective wellbeing tend also to exhibit other cues which we associate with higher wellbeing: they smile more during interviews, they are rated as happier by friends and relatives, they exhibit lower physical manifestations of stress, and the parts of their

³ Subjective wellbeing is defined here as self-reported responses to survey questions probing any aspect of wellbeing, though most commonly happiness and satisfaction. The term wellbeing is not necessarily synonymous with utility, which might be considered more specific and context-dependent in some applications. As discussed further on, available evidence suggests subjective wellbeing is likely to be a good proxy for experience utility, though not for other types of utility like decision utility. Consequently, in the discussions provided here utility refers specifically to experience utility. Likewise, happiness and satisfaction, which are two common types of subjective wellbeing measures, are not necessarily synonymous. Crooker and Near (1998) provides a discussion on the respective meanings of happiness and satisfaction, but this is not considered in further detail here.

⁴ In other words, such measurement scales exhibit interval-level quality. A possible further assumption implies ratio-level quality, which requires a unique and non-arbitrary zero-point (or point of neutrality). This level of quality is not discussed directly here, as it is considered quite contentious and to have limited consequences for common uses of subjective wellbeing data. The case for ratio quality is discussed in some detail in Kristoffersen (2010).

⁵ Thus, the condition of uniqueness would be unnecessary in fixed- and random-effects panel models, because individual fixed effects are accounted for. Hirschauer et al. (2014) elaborate more specifically on the conditions for comparison across individuals, within individuals over time, and within individual across categories.

brains associated with pleasant emotions are more activated.⁶ Clark et al. (2008) evaluate the evidence for assuming subjective wellbeing data also capture relevant information about utility, and conclude that common survey-based measures seem highly likely to reflect experience utility, and are therefore appropriate for welfare analysis in many cases.⁷ Consequently, this discussion proceeds under the assumption that subjective wellbeing is a valid measure of psychological wellbeing and experience utility.

2.2.2 Ordinal Comparability

Ordinal comparability may seem a reasonable assumption within individuals: If Jack reports a subjective wellbeing score of 5 yesterday and 6 today, then it seems reasonable to assume that his wellbeing, and utility, has increased. However, it is perhaps less certain that Jill, who scores 7 today, is in fact experiencing a greater true wellbeing or utility than Jack. We are therefore faced with the common problem of arbitrary and ambiguous measurement (Blanton and Jaccard 2006). Even though positive monotonic relationships are observed between reported subjective wellbeing and other observable cues of psychological wellbeing, it is far from clear that people who score at different points on the measurement scale really are different, in the way we expect. It is also possible that the length of some scales, particularly the eleven-point measurement scale used in the HILDA survey and others, are longer than what is commonly considered ideal in terms of individuals' ability to make distinct judgements, which is around seven points (Miller 1956).

2.2.3 Cardinal Comparability

Cardinal comparability of subjective wellbeing scores implies that the difference between score points of 5 and 6, in terms of wellbeing or utility, is equal to the difference between 6 and 7, 7 and 8, and so forth. As mentioned, numeric measurement scales convey some intention of cardinality, and research into the perceptions of these survey instruments for the measurement of subjective wellbeing have revealed that people interpret them as cardinal, and intend to provide responses that reflect this as accurately as possible (Van Praag 1991; Parducci 1995; Schwartz 1995). This would suggest that the psychological concepts of wellbeing and utility are themselves cardinal, which is a necessary condition for cardinal measurement. Most latent psychological concepts, such as intelligence, are treated as cardinal, and the analysis therefore proceeds under the assumption that this applies also to true wellbeing and utility. Ng (1996) argues that

⁶ See for example Larsen and Fredrickson (1999), Van Praag (1991), Sandvik et al. (1993), and Diener and Lucas (1999). Brief surveys of the collective evidence are provided by Clark et al. (2008) and by Layard et al. (2008).

⁷ First, implicit trade-offs, as measured in empirical models of subjective wellbeing, generally correspond well with what we know about choice behaviour: for example, the observed positive effects of marriage and employment on subjective wellbeing correspond well with the amount of effort people tend to put into obtaining these outcomes. Second, observed behaviour is consistent with what we expect from wellbeing-maximising individuals: for example, low satisfaction scores in the spheres of work and marriage tend to be good predictors of job change and divorce. Finally, the evidence which emerges from the analysis of survey data on subjective wellbeing corresponds well with that which emerges from experimental economics, particularly with respect to positional concerns (Clark et al. 2008).

individuals are able, when provided with the necessary tools, to provide data which are not only of interval but also of ratio quality.⁸

2.2.4 Sources of Nonlinearity in Measurement Scales

Survey instruments will necessarily impose some restrictions at the edges of the measurement scale by forcing a variable which presumably is unbounded onto a bounded scale, as argued by Ng (2008). This implies a logistic response function for subjective wellbeing, with distances between score-points on the measurement scale increasing toward both extremes of the scale. Alternatively, bounded utility may be a justifiable assumption, since marginal utility approaches zero as people approach the point of satiation.⁹ If the measurement scale is bounded in such a way that it approximates utility, and does not impose noticeable restrictions, the response function for that measure will be approximately linear. Consequently, under certain assumptions, subjective wellbeing might be considered an acceptable approximation of utility (Hirschauer et al. 2014).

When comparing across individuals we face a potential source of ambiguity if individuals differ in their attitudes toward scoring at the extremes of the measurement scale.¹⁰ This will make differences at the edges of the scale more ambiguous and less distinct, rather than more distinct. The possibility of ambiguity and nonlinearity of the measurement scale is therefore difficult to dismiss without any further information, also when adopting a fully operationalised definition of subjective wellbeing.

This description of a logistic relationship between true and observed subjective wellbeing closely resembles that between stimulus and response, which forms the basis of fundamental utility theory and the idea of diminishing marginal utility. Specifically, the Weber-Fechner law holds that the relationship between stimuli and sensation is logarithmic [sensation = $k \ln(\text{stimulus})$] (Masin et al. 2009), which implies a logistic relationship if the underlying concept is bipolar.¹¹ Consequently, the shape of the utility function, which maps the relationship between life circumstances (such as consumption, income and wealth) and utility, is often assumed to be logarithmic. While potentially related, this stimulus–response mechanism is treated here as distinct from the measurement issue of how true subjective wellbeing translates into reported subjective wellbeing, which is the main focus of this paper. Hence, these relationships are hypothesised to potentially take the same functional form, though for slightly different reasons.¹²

The possibilities described above produce a set of hypothesised functional forms for the response function which translates true wellbeing or utility (u) into observed survey

⁸ Ng (1996) uses a small but complex survey, of 41 Chinese graduate students, that individuals are capable of making judgements on how much time they spend in unhappy and happy states, and agree with the notion of a scale of net happiness with a unique zero-point. While this evidence is of value, Ng himself admits the sample is clearly small and selective, and that the questions are slightly leading in nature.

⁹ Similarly, utility is assumed to be bounded in von Neuman–Morgenstern utility models (Savage 1954).

¹⁰ Some evidence of such effect in subjective wellbeing data are provided by Lau (2007).

¹¹ The specific idea of a logarithmic utility function appears first to have been proposed by Bernoulli (1738 [1954]). Edgeworth (1881 [1961]), who proposed the idea of a ‘hedonimeter’ which would produce an objective and fully cardinal measure of utility, was similarly interested in the idea of just perceivable increments.

¹² Ng (1996, 2008) suggests increasing increments of true happiness (which is considered a stimulus in that context) are required to produce perceivable increments in measured happiness (the response). Hence, Ng’s exposition of a logistic response function incorporates elements of the stimulus–response relationship as well the possibility of bias implied by the boundedness of measurement scales.

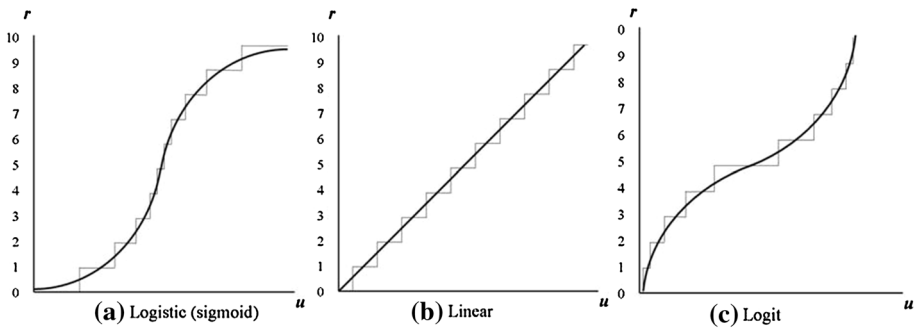


Fig. 1 Hypothesized response functions. *Note* The stepped lines in this diagram reflect discrete measurement scales. **a** Is adapted from Ng (2008). **b** The notation used are sourced from Blanchflower and Oswald (2004)

responses (r), as illustrated in Fig. 1.¹³ If bias from scale restrictions dominate, the response function is hypothesised to be shaped like a logistic function (S -shaped), as shown in graph (a). If the measurement scale approximates the boundedness of true wellbeing or utility, the response function will be linear, as shown in graph (b). If end-of-scale ambiguities dominate, then the response function is hypothesised to be shaped like a logit function (inverse S -shaped), as shown in in graph (c).

2.3 Problems of Unobservable Variables and Arbitrary Scales

Utility and true wellbeing are latent variables which remain unobservable. Therefore, any attempt at evaluating cardinal comparability of subjective wellbeing data may seem futile, unless some operationalised definition which circumvents this problem can be justified and adopted. However, according to fundamental measurement theory this problem can never be completely avoided, because all measures are inferences by definition (Wright 1997). In psychometrics, the problem of measuring abstract latent concepts is approached by a form of operationalised measurement. For example, a measure for ability is constructed by, first, defining what this type of ability entails, and second, designing questions or tasks that will generate the necessary data to capture this information.

Such measurement instruments are commonly designed to fit particular logistic probability distributions as specified by Rasch models (originating in Rasch 1961). The ‘operationalised’ response functions for these measures are therefore logistic in shape (S -shaped) in the same way as described by Ng (2008), and essentially for similar reasons. As the exact curvature of the response function is given by the Rasch model, raw scores are transformed as required in order to enable cardinal comparison (Wright 1997).¹⁴ Consequently, measures which are constructed and verified to fit the relevant Rasch model can be

¹³ In this figure, a zero-point or point of neutrality of utility or wellbeing might reasonably be inferred by the inflection points of panels (a) and (c), which may or may not correspond with the mid-point of the measurement scale. This would incorporate the idea of negative utility or wellbeing, which is reasonable where utility or wellbeing is perceived as the net of positive and negative feelings or experiences, as suggested by Ng (1996).

¹⁴ Specifically, Rasch models apply additive conjoint measurement (Luce and Tukey 1964) to produce a measure where conjoint transitivity implies that items and persons are measured on an interval scale with a common unit (Brogden 1977; Wright 1997). Andrich (1978) later developed the Polytomous Rasch model

interpreted as having a known response function, which is linear when adjusted scores are used. While the intention is to produce a cardinal measure, one might argue that this type of response function describes the statistical features of a particular measure, and is not equivalent to the function which translates the true latent concept into observed responses.

Subjective wellbeing measures are different from these types of psychometric measures in two important ways. Firstly, they are measured by single-item Likert scales rather than multiple responses. Consequently, without additional points of reference we know very little about these measurement scales compared to other measures based on multiple items (questions). Secondly, they differ from other such instruments in the sense that the definition of happiness or satisfaction—what it means to be happy or satisfied—rests with the respondent rather than the researcher. Consequently, subjective wellbeing measures are more ambiguous and arbitrary in terms of uniqueness and comparability—particularly across individuals.

Blanton and Jaccard (2006) consider the problem of arbitrary and ambiguous satisfaction scales (and other similar Likert scales), and propose that we address this problem by obtaining additional information to provide some reference points to qualify and quantify the gaps of the measurement scale using meaningful information. For example, in the case of subjective wellbeing one might measure smiling frequency during interviews, cortisol levels in the blood, or activity in the area of the brain associated with pleasure. This provides a means of evaluating ordinal comparability (uniqueness) of subjective wellbeing scales, and potentially also of evaluating cardinal comparability (equidistance).¹⁵

The premise of the analysis to follow is that, although we may not be able to observe utility or true wellbeing and the response function for subjective wellbeing directly, the principle of simultaneous conjoint measurement provides a means of indirectly observing something about its possible shape. Similarly, in order to derive full information about the true shape of a three-dimensional object we must be able to observe it from all possible angles, though this is not possible with latent abstract concepts. Observing the object from only one single angle will provide very limited information about its true shape, and subsequent angles add information of potentially significant (and diminishing) value.

Similarly, this paper seeks to improve our understanding of the eleven-point numeric life satisfaction scale by comparing people who score at different points on this scale according to available information on mental health. Because these mental health data are constructed to fit the relevant Rasch model we have some, potentially quite accurate, information about the shape of the response function for mental health, and we may thereby be able to evaluate the possible shape of the response function for life satisfaction. The method of analysis is explained in further detail in the following section.

Footnote 14 continued

for multiple ordered (rather than dichotomous) responses. See Wright's (1997) for a brief description of the history and development of measurement in social sciences.

¹⁵ Thus, the solution proposed by Blandon and Jaccard (2006) applies the same principles of simultaneous conjoint measurement upon which Rasch models are founded. As explained by Luce and Tukey (1964): "the essential character of simultaneous conjoint measurement is described by an axiomatization of the comparison of effects of (or responses to) *pairs* formed from *two specific kinds* of 'quantities'." They explain that these can potentially produce a cardinal (interval-quality) measure: "The axioms apply when, for example, the effect of a pair consisting of one mass and one difference in gravitational potential on a device that responds to momentum is compared with the effect of another such pair. Measurement on interval scales which have a common unit follows from these axioms".

3 Method of Analysis: Evaluating the Response Function for Subjective Wellbeing

3.1 Combining Two Response Functions

Following the notation used in Blanchflower and Oswald (2004), the relationship between true wellbeing and the observed response can be modeled as follows:

$$r = h(u) + e. \quad (1)$$

Here, r is the individual's reported wellbeing score, u is to be interpreted as the individual's true (unobservable) wellbeing or utility, h is the function that transforms true wellbeing into reported wellbeing, and e is an error term. Following the discussion above, the shape of function h is hypothesised to lie somewhere in the space between logistic and logit, including linearity, as illustrated earlier in Fig. 1.

Information about wellbeing or utility may be captured by a range of different types of measures aside from conventional subjective wellbeing measures. Underlying each such measure is a response function which captures information about the same unobservable concept into a particular instrument. That is, u translates into r (reported subjective wellbeing) via the function h , and also into some other alternative wellbeing measure we can call w via another function we can call g . Since u is unobservable both h and g are also unobservable. However, using the rule of transitivity, certain features of h and g may be observed indirectly via a third function we can call k , which describes the relationship between r and w . Formally (ignoring the error terms):

$$r = h(u) \quad w = g(u) \rightarrow r = h[g^{-1}(w)] = k(w) \quad (2)$$

This means that the shape of the observable function k is a result of the combination of h and the inverse of g . The observed form of k then implies a potentially limited set of possibilities with respect to the shapes of functions h and g . In particular, if we know something about the shape of g , and function k is reliably estimated, the range of possible shapes of h may be quite narrow.

The set of possible response functions produces a limited set of possible shapes of the observable function k , which itself also likely lies somewhere in the spectrum between logistic and logit, depending on the shape of function g . For example, a linear function k can only result from functions g and h taking exactly the same form (with the same strength in curvature). This is because function k is function h transformed by g^{-1} , so if h and g have the same shape and curvature function k will be linear. If g and h take opposite forms, then the form of k will be an exaggeration of h . Of course, many other possibilities exist, as summarised in Table 1.¹⁶

If the mental health measure used here can be assumed to be linearly (or approximately linearly) related to true wellbeing or utility, then the shape of the response function for life satisfaction would be observed indirectly, via the relationship between life satisfaction and mental health. A linear relationship between mental health and life satisfaction then implies a linear response function for life satisfaction, in terms of true wellbeing or utility.

¹⁶ If function k is found to be irregular we could either conclude that no recognisable pattern exists, and that linearity is not a reasonable assumption, or the range of possible functional forms could potentially be expanded in search of a recognisable pattern and a functional form that enables transformation of subjective wellbeing data onto a linear scale.

Table 1 The possible shapes of function k , given the shapes of functions h and g

	Function h				
	S	s	LIN	is	IS
Function g					
S	<i>LIN</i>	<i>is</i>	<i>IS</i>	<i>IS*</i>	<i>IS**</i>
s	<i>s</i>	<i>LIN</i>	<i>is</i>	<i>IS</i>	<i>IS*</i>
LIN	<i>S</i>	<i>s</i>	<i>LIN</i>	<i>is</i>	<i>IS</i>
is	<i>S*</i>	<i>S</i>	<i>s</i>	<i>LIN</i>	<i>is</i>
IS	<i>S**</i>	<i>S*</i>	<i>S</i>	<i>s</i>	<i>LIN</i>

Response function type and curvature are indicated as follows: *LIN*, linear; *s*, weakly S-shaped (logistic); *S*, moderate S-shaped (logistic); *S**, strong S-shaped (etc.); *is*, weak inverse S-shaped (logit); *IS*, moderate inverse S-shaped (logit); *IS**, strong inverse S-shaped (etc.). Definitions of what constitutes weak, moderate and strong curvatures are not given. The point here is to show that if both response functions g and h have the same functional form, then the combined function k will be perfectly linear, while if g and h have opposite curvatures the function k will be either logistic shaped or logit shaped, as indicated

In general, the shape of the relationship between mental health and life satisfaction determines the possible shapes that the response function for life satisfaction can take. If we have no a priori knowledge about the possible shapes of the response functions for mental health or life satisfaction, then anything is possible. However, if we know something about either, we can narrow down the possibilities.

3.2 The Response Function for the MH5 Mental Health Index

The measure of mental health used here is the MH5 index, which in its raw form is a five-item aggregate score and part of the SF36 survey instrument for measuring health. This index is generated by asking the question ‘How much of the time during the past 4 weeks (a) have you been a nervous person, (b) have you felt so down in the dumps that nothing could cheer you up, (c) have you felt calm and peaceful, (d) have you felt down, and (e) have you been a happy person’. Responses are coded to a six-point scale of (1) all of the time, (2) most of the time, (3) a good bit of the time, (4) some of the time, (5) a little of the time, and (6) none of the time. The raw MH5 score is calculated by first reversing the scores where appropriate such that higher values indicate better mental health, then adding the score for each question, and finally standardising this sum to a 0–100 index (Ware et al. 2000).

The MH5 index has been constructed such that observed responses conform to the probabilistic features of the Rasch model, and scores have subsequently also been found to fit this model very well (Perneger and Bovier 2001), as has all main components of the SF36 health instrument (Raczek et al. 1998). Consequently, raw MH5 scores must be transformed or adjusted to eliminate the ‘raw score bias’ such that the data can be treated as cardinal, as described by Brogden (1977).¹⁷ This may then be interpreted to specifically

¹⁷ Specifically, the raw MH5 index scores intervals 0–10, 10–20, etc., up to 90–100 have logit intervals of 2.23, 1.22, 1.00, 0.90, 0.85, 0.83, 0.90, 1.09, 1.54, and 3.21 (Perneger and Bovier 2001). Accordingly, the following transformation function will linearise these intervals: $MH5^T = \ln\left(\frac{0.00932MH5+0.034}{1-(0.00932MH5+0.034)}\right)$. This produces a scale with lower and upper bounds of -3.35 and $+3.35$, with a mid-point of zero. For convenience, this is scaled to produce a 0–100 index in the analysis to follow.

imply a linear response function for adjusted MH5 scores, or to provide some relevant but approximate information about of the response function for raw MH5 scores, or to be irrelevant.

3.3 The Association Between Life Satisfaction and Mental Health

The possible shape of the response function for life satisfaction is here observed indirectly through the association between life satisfaction and raw or adjusted (cardinalised) MH5 scores. As a starting point, the presence of nonlinearities in this association is statistically evaluated by estimating life satisfaction (LS) as a function of mental health ($MH5$), with linear, squared and cubed terms included (Eq. 3).

$$LS_i = \gamma_0 + \gamma_1(MH5_i) + \gamma_2(MH5_i)^2 + \gamma_3(MH5_i)^3 + u_i \quad (3)$$

In this case, the relationship between life satisfaction and the mental health index is hypothesised to fall somewhere in the spectrum between logistic and logit in shape. The extent to which the data fit these specific shapes is evaluated by estimating Eqs. (4) and (5), which represent standard logistic and logit functions, respectively.

$$LS_i = \frac{\lambda}{1 + e^{-(\alpha + \beta MH5_i)}} + v_i \quad (4)$$

$$e^{LS_i} = \frac{\alpha + \beta(MH5_i)}{1 - \alpha - \beta(MH5_i)} + v_i \quad (5)$$

Ordinal and cardinal interpersonal comparability of life satisfaction scores, in terms of MH5, is specifically evaluated by comparing the MH5 scores of people who score at different points of the eleven-point life satisfaction scale. This is facilitated by estimating Eq. (6), where raw and adjusted mental health scores are regressed on a set of dummy variables indicating the life satisfaction group to which each sampled individual belongs.¹⁸

$$MH5_i = \beta_0(LS_{0,i}) + \beta_1(LS_{1,i}) + \dots + \beta_9(LS_{9,i}) + \beta_{10} + \varepsilon_i. \quad (6)$$

LS_0 takes the value 1 for individuals with a life satisfaction score of 0, and a value of 0 otherwise, and so forth. Individuals who report a life satisfaction score of 10 form the control group. The intercept term β_{10} will therefore return the mean mental health value for this group, and the other β 's measure the distance, in mean mental health scores, between each respective LS group and the control group.¹⁹ The parameter ε is an error term. The differences or shifts in mean mental health scores across life satisfaction groups are evaluated visually, to determine how the data behave with respect to the hypothesized shapes illustrated in Fig. 1; and statistically, to determine ordinal distinctness and equidistance.

Ordinal distinctness (interpersonal comparability) of life satisfaction scores, with respect to mental health information, is evaluated by testing that the β 's in Eq. (6) are

¹⁸ Measurement theory posits that raw such scores are not to be used, as they are not true measures (that is, they are not linear) (Wright 1997). However, the results presented here are generated using both raw and adjusted scores in the interest of completeness, as this provides better information about the implications of the adjustment.

¹⁹ Calculating mean mental health scores for each life satisfaction group will produce the same information, but by estimating differences in means in this way we are also able to perform the necessary hypothesis tests of model parameters.

different and follow the expected order (hypothesis $H1$). Equidistance of score differences is evaluated by testing that these β 's increase with equal increments (hypothesis $H2$).

$$H1: \beta_0 < \beta_1 < \dots < \beta_8 < \beta_9 \tag{7}$$

$$H2: (\beta_1 - \beta_0) = (\beta_2 - \beta_1) = \dots = (\beta_9 - \beta_8) = -\beta_9 \tag{8}$$

If the data reject $H1$, life satisfaction scores cannot be considered ordinally distinct in terms of mental health information. Cardinal comparability is therefore also rejected by default. If the data do not reject $H1$, ordinal comparability across individuals, in terms of MH5 data, is confirmed. If neither $H1$ nor $H2$ are rejected, then linearity of the function k is confirmed statistically. This may then be considered a justifiable basis for assuming cardinal comparability across individuals, though this conclusion rests on the assumption that (1) the comparison between mental health and life satisfaction is meaningful (i.e. that the function k is estimated with reasonable fit), and (2) the response function for mental health, g , is linear.

4 Data and Analysis

4.1 Core Analysis: The Interpersonal Association Between Life Satisfaction and Mental Health

Table 2 presents summary statistics of the life satisfaction and mental health data from waves 1 to 11 of the HILDA survey. Mental health scores (both raw and adjusted) increase

Table 2 Descriptive statistics of the mental health and life satisfaction data

Raw mental health ($MH5, w$)			Life satisfaction (LS, r)		
<i>N</i>	129,847		<i>N</i>	129,847	
<i>Individuals</i>	24,082		<i>Individuals</i>	24,082	
<i>Mean</i>	74.26		<i>Mean</i>	7.94	
<i>Median</i>	80		<i>Median</i>	8	
<i>SD</i>	17.06		<i>SD</i>	1.48	
Score interval	Frequency (%)	Score groups (j)	Frequency (%)	Mean raw $MH5 \bar{w}_j$	Mean adj. $MH5 \bar{w}_j^*$
		0	0.12	40.06	42.83
0–9	0.21	1	0.17	35.15	39.4
10–19	0.38	2	0.37	42.22	44.7
20–29	1.39	3	0.71	45.14	46.59
30–39	1.83	4	1.15	51.11	50.63
40–49	5.79	5	4.07	56.23	53.98
50–59	7.74	6	5.80	62.22	57.83
60–69	16.35	7	18.72	69.52	62.72
70–79	16.12	8	33.48	75.98	67.56
80–89	33.71	9	21.93	80.56	71.49
90–100	16.79	10	13.46	82.17	74.04

This sample consists of all individuals aged 15 and above for whom scores on life satisfaction and mental health is available, from waves 1 to 11 of the HILDA survey

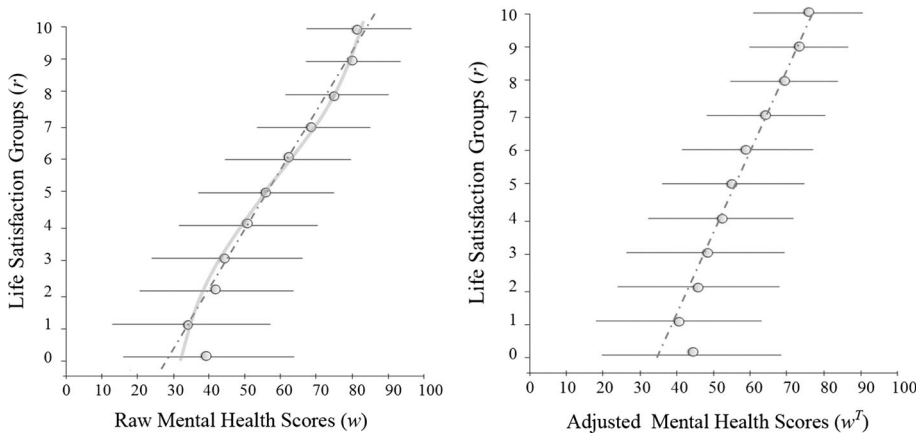


Fig. 2 Mental health characteristics of life satisfaction groups: across individuals

monotonically across the life satisfaction scale, and follow the expected order. The lowest life satisfaction group (those who score zero) represents an outlier, though this group is also very small. The movements in raw mean mental health scores across each life satisfaction group fluctuate somewhat around an average of about 5 (weighted by the number of observations).

Figure 2 displays mean raw and adjusted mental health scores of each life satisfaction group, with a line extending one standard deviation in each direction. The broken straight lines superimposed upon these diagrams represent the best-fit linear regression line. The first diagram gives a strong impression of a near-linear association between raw mental health and life satisfaction, though with a weak tendency for score distances to diminish toward the upper edge of the scale—i.e. a weakly logit-shaped functional form. This is illustrated by the solid shaded curve superimposed upon the diagram. When transformed mental health data are used this nonlinearity seems to disappear, producing an association which appears approximately linear.

Estimates of Eqs. (3), (4) and (5) are presented in Tables 3 and 4. Nonlinearities are captured both when raw and adjusted mental health scores are used, though these are much weaker when adjusted data are used. Moreover, the curvatures of these two functions are different, indicating a logit (inverse S-shaped) function when using raw mental health data and potentially a weakly logistic (S-shaped) function when using adjusted data. The weak logit shape of the function k observed when using raw mental health data is confirmed when this function is estimated using Eq. (5). When using adjusted mental health data the logit function is rejected, but significant parameters for the logistic function (Eq. 4) is estimated instead, suggesting a weak logistic shape is captured in this association.

Estimates of Eq. (6) are provided in Table 5, along with test statistics for ordinal distinctness (hypothesis $H1$) and equidistance of score-points ($H2$), in terms of raw and adjusted mental health data. Hypothesis $H1$ is effectively tested by observing that the parameters have the correct order and then testing the alternative hypothesis that these parameters are equal. This alternative hypothesis is rejected for both raw and adjusted mental health data by joint significance tests (though only when the LS_0 group is omitted, since the order condition is violated for this group). Hypothesis $H2$ is rejected in both cases. Pairwise tests of parameters are also provided, where the distance between life

Table 3 Life satisfaction as a cubic function of mental health scores

<i>Parameter</i>	Raw MH5 scores <i>Coefficient (SE)</i>	Adjusted MH5 scores <i>Coefficient (SE)</i>
<i>Intercept</i> (γ_0)	3.74*** (0.0659)	3.15*** (0.0994)
<i>Linear term</i> (γ_1)	0.11*** (0.0036)	0.08*** (0.0048)
<i>Squared term</i> (γ_2)	-0.0012*** (0.000062)	0.0000 (0.000079)
<i>Cubed term</i> (γ_3)	0.0000064*** (0.00000034)	-0.0000022*** (0.00000042)
<i>Adjusted R²</i>	0.2134	0.2128
<i>F-statistic</i>	11,746.13***	11,704.62***
<i>N</i>	129,847	129,847

These estimates are based on pooled data from waves 1–11 of the HILDA survey. The sample is restricted to people aged 15 and above. MH5 denotes mental health scores, scaled to a 0–100 index. Statistical significance at the 90, 95 and 99 % level of confidence is indicated by *, **, and ***, respectively

Table 4 Life satisfaction as a logistic/logit function of mental health scores

<i>Parameter</i>	Raw MH5 scores: logit function <i>Coefficient (SE)</i>	Adjusted MH5 scores: logistic function <i>Coefficient (SE)</i>
α	0.9220*** (0.0005)	-0.7545*** (0.0161)
β	0.00067*** (0.000006)	0.0338*** (0.0006)
λ	–	9.82*** (0.0421)
<i>Adjusted R²</i>	(0.4959)	(0.9735)
<i>F-statistic</i>	Not available	Not available
<i>N</i>	129,847	129,847

These estimates are based on pooled data from waves 1–11 of the HILDA survey. The sample is restricted to people aged 15 and above. MH5 denotes mental health scores, scaled to a 0–100 index. Note that the adjusted *R*-squared for the nonlinear regression is not interpreted in the normal way because no constant term is included. Statistical significance at the 90, 95 and 99 % level of confidence is indicated by *, **, and ***, respectively

satisfaction scores of 4 and 5 is chosen as the representative distance with which the other distances are compared, as this is closest to the weighted mean of score distances. The distance between the two lowest life satisfaction groups is assessed against the hypothesis that this distance is zero (rather than negative compared to what is expected). Half of the pairwise tests are rejected, and the other half are not.

These results imply that people who select different life satisfaction scores are distinct in terms of their reported mental health, regardless of whether raw or adjusted data are used. Furthermore, the mental health scores of these people follow the expected order (except for people with a life satisfaction score of zero, who are clearly somewhat different). Therefore, the assumption of ordinal distinctness of life satisfaction scores across individuals, in terms of mental health information, is supported by these data.

The association between life satisfaction scores and raw mental health data is logit-shaped with a fairly weak curvature, with distances between score-points diminishing slightly toward the upper end of the life satisfaction scale. This observed nonlinearity is corrected for when adjusted mental health data are used, producing an association which is

Table 5 Raw and adjusted mental health across life satisfaction groups: linear regression estimates

Parameters	Raw MH5 (SE)	Adjusted MH5 (SE)
<i>Intercept</i> ($\beta_{10} = \bar{w}_{10}$)	82.17*** (0.1139)	74.04*** (0.0889)
$\beta_9 (= \bar{w}_9 - \bar{w}_{10})$	-1.61*** (0.1448)	-2.55*** (0.1130)
$\beta_8 (= \bar{w}_8 - \bar{w}_{10})$	-6.19*** (0.1349)	-6.48*** (0.1053)
$\beta_7 (= \bar{w}_7 - \bar{w}_{10})$	-12.64*** (0.1494)	-11.32*** (0.1166)
$\beta_6 (= \bar{w}_6 - \bar{w}_{10})$	-19.95*** (0.2076)	-16.21*** (0.1620)
$\beta_5 (= \bar{w}_5 - \bar{w}_{10})$	-25.94*** (0.2365)	-20.06*** (0.1845)
$\beta_4 (= \bar{w}_8 - \bar{w}_{10})$	-31.05*** (0.4054)	-23.41*** (0.3164)
$\beta_3 (= \bar{w}_3 - \bar{w}_{10})$	-37.02*** (0.5104)	-27.45*** (0.3983)
$\beta_2 (= \bar{w}_2 - \bar{w}_{10})$	-39.95*** (0.5935)	-29.34*** (0.5413)
$\beta_1 (= \bar{w}_1 - \bar{w}_{10})$	-47.02*** (1.0130)	-34.64*** (0.7906)
$\beta_0 (= \bar{w}_0 - \bar{w}_{10})$	-42.01*** (1.2039)	-31.21*** (0.9396)
<i>Model information and diagnostics</i>	$N = 129,847$	$N = 129,847$
	$\bar{R}^2 = 0.2200$	$\bar{R}^2 = 0.2077$
	$F = 3663.07***$ ($p < 0.0000$)	$F = 3405.54***$ ($p < 0.0000$)
<i>Joint hypothesis test: $\beta_1 = \beta_2 = \dots = \beta_8 = \beta_9$</i>	$F = 3451.53***$ ($p < 0.0000$)	$F = 2923.97***$ ($p < 0.0000$)
<i>Joint hypothesis test:</i> $(\beta_1 - \beta_0) = (\beta_2 - \beta_1) = \dots = (\beta_9 - \beta_8) = -\beta_9$	$F = 166.28***$ ($p < 0.0000$)	$F = 52.45***$ ($p < 0.0000$)
<i>Pairwise hypothesis tests: $(-\beta_9) = (\beta_5 - \beta_4)$</i>	$F = 57.17***$ ($p < 0.0000$)	$F = 4.79**$ ($p = 0.0286$)
$(\beta_9 - \beta_8) = (\beta_5 - \beta_4)$	$F = 1.36$ ($p = 0.2439$)	$F = 2.65$ ($p = 0.1036$)
$(\beta_8 - \beta_7) = (\beta_5 - \beta_4)$	$F = 8.55***$ ($p = 0.0035$)	$F = 17.39***$ ($p < 0.000$)
$(\beta_7 - \beta_6) = (\beta_5 - \beta_4)$	$F = 20.48***$ ($p < 0.0000$)	$F = 16.87***$ ($p < 0.0012$)
$(\beta_6 - \beta_5) = (\beta_5 - \beta_4)$	$F = 2.16$ ($p = 0.1414$)	$F = 1.16$ ($p = 0.2820$)
$(\beta_4 - \beta_3) = (\beta_5 - \beta_4)$	$F = 0.82$ ($p = 0.3657$)	$F = 0.87$ ($p = 0.3505$)
$(\beta_3 - \beta_2) = (\beta_5 - \beta_4)$	$F = 5.28**$ ($p = 0.0216$)	$F = 3.79*$ ($p = 0.0515$)
$(\beta_2 - \beta_1) = (\beta_5 - \beta_4)$	$F = 2.27$ ($p = 0.1317$)	$F = 3.72*$ ($p = 0.0539$)
$(\beta_1 - \beta_0) = 0$	$F = 9.82***$ ($p = 0.0017$)	$F = 7.86***$ ($p = 0.0051$)

These estimates are based on pooled data from waves 1–11 of the HILDA survey. The sample is restricted to people aged 15 and above. Statistical significance at the 90, 95 and 99 % level of confidence is indicated by *, **, and ***, respectively

visually very close to linear. However, the adjustment of mental health scores appears to ‘overcompensate’, and equidistance is strictly rejected by the statistical tests performed here.

The model estimates presented here suggest information about mental health explain around 21 % of the variation in life satisfaction observed across individuals. The degree of

commonality of information between these variables may therefore seem fairly modest, though this level of explanatory power exceeds that of many large subjective wellbeing models presented in the literature.²⁰

4.2 An Extension: Intertemporal Comparisons of Life Satisfaction and Mental Health

Subjective wellbeing data are increasingly available in longitudinal panels. Intrapersonal, or intertemporal, comparison is therefore of interest, in addition to interpersonal comparison. It is difficult to evaluate intertemporal comparison with the same level of specificity as interpersonal comparison. Nonetheless, the analysis is extended to consider intertemporal comparison in as far as this is possible, providing some supplementary results to evaluate whether the patterns observed across individuals is also observed within individuals across time. As such, Eq. (6) is estimated as a fixed-effects panel model, with results presented in Table 6, and visually in Fig. 3.

A very similar pattern is observed within individual across time as across individuals. Ordered distinctness of score-points is confirmed by a joint hypothesis test, apart from the lowest score which again is an anomaly. Movements in raw mental health scores across life satisfaction score-points are again observed to diminish slightly toward the upper end of the life satisfaction scale. When adjusted mental health scores are used these nonlinearities are again diminished, and the relationship appears approximately linear, although the condition of equidistance is again rejected by the joint hypothesis test.

5 Summary, Discussion and Conclusion

The increasingly common use of survey data on happiness and satisfaction in economic analysis and for policy design and evaluation implies a degree of acceptance that these data bear a meaningful relationship with the relevant types of utility. The general assumption of a positive monotonic relationship between utility and subjective wellbeing is common, and seems justified, though these data are also increasingly assumed to be cardinally comparable both across and within individuals. Hirschauer et al. (2014, p. 654) state that “Aside from the mismatch between utility and happiness, the behavior of the measurement function itself [...] is far from clear. Due to its strong focus on empirical research, this is often overlooked in happiness research. As a result, cardinal interpretability and interpersonal comparability of subjective well-being data are often taken as a given.” This paper addresses the issues of ordinal and cardinal comparability of these data directly.

Ordinal comparability of subjective wellbeing scores is implicitly assumed in most (or all) analyses of such data, though there is very little specific evidence on which to base

²⁰ When these life satisfaction data are regressed on gender, age (squared and cubed terms included), marital status, the presence of children, labour force participation, physical health, education, income (log of equalised household income) and personal characteristics, these variables are found to explain just over 18 % of the variation in life satisfaction observed across individuals. This is comparable to other similar models presented in the literature, such as Boyce and Wood (2011), who estimate a similar model with an adjusted R-squared of 0.20. When mental health information is added to such a model the explanatory power increases from 18 to over 27 %. Consequently, mental health information contains a lot of information about life satisfaction, much of which is unique and cannot be attributed to other variables. Because this type of life satisfaction model is not of key interest here, these estimates are not presented in the paper.

Table 6 Raw and adjusted mental health within individuals: fixed-effects panel model estimates

Parameters	Raw MH5 (SE)	Adjusted MH5
<i>Intercept</i> ($\beta_{10} = \bar{w}_{10}$)	78.73*** (0.1133)	70.49*** (0.0879)
$\beta_9 (= \bar{w}_9 - \bar{w}_{10})$	-1.27*** (0.1337)	-1.40*** (0.1038)
$\beta_8 (= \bar{w}_8 - \bar{w}_{10})$	-3.57*** (0.1355)	-3.27*** (0.1052)
$\beta_7 (= \bar{w}_7 - \bar{w}_{10})$	-6.98*** (0.1336)	-5.74*** (0.1192)
$\beta_6 (= \bar{w}_6 - \bar{w}_{10})$	-10.81*** (0.1975)	-8.26*** (0.1534)
$\beta_5 (= \bar{w}_5 - \bar{w}_{10})$	-13.73*** (0.2230)	-10.06*** (0.1731)
$\beta_4 (= \bar{w}_8 - \bar{w}_{10})$	-17.25*** (0.3525)	-12.33*** (0.2736)
$\beta_3 (= \bar{w}_3 - \bar{w}_{10})$	-20.64*** (0.4379)	-14.74*** (0.3399)
$\beta_2 (= \bar{w}_2 - \bar{w}_{10})$	-22.37*** (0.5881)	-16.02*** (0.4566)
$\beta_1 (= \bar{w}_1 - \bar{w}_{10})$	-27.83*** (0.8524)	-20.34*** (0.6617)
$\beta_0 (= \bar{w}_0 - \bar{w}_{10})$	-21.72*** (1.0306)	-15.94*** (0.8001)
<i>Model information and diagnostics</i>	$N = 129,847$ (24,082)	$N = 129,847$ (24,082)
	$\bar{R}^2 = 0.6108^\#$	$\bar{R}^2 = 0.6090^\#$
	$F = 878.94^{***}$ ($p < 0.0000$)	$F = 750.45^{***}$ ($p < 0.0000$)
<i>Joint hypothesis test: $\beta_1 = \beta_2 = \dots = \beta_8 = \beta_9$</i>	$F = 884.31^{***}$ ($p < 0.0000$)	$F = 715.56^{***}$ ($p < 0.0000$)
<i>Joint hypothesis test:</i> $(\beta_1 - \beta_0) = (\beta_2 - \beta_1) = \dots = (\beta_9 - \beta_8) = -\beta_9$	$F = 35.04^{***}$ ($p < 0.0000$)	$F = 17.54^{***}$ ($p < 0.0000$)

These estimates are based on pooled data from waves 1–11 of the HILDA survey. The sample is restricted to people aged 15 and above. Statistical significance at the 90, 95 and 99 % level of confidence is indicated by *, **, and ***, respectively

This statistic is generated by estimating the model with a dummy variable included for each individual surveyed in the sample. This produces an R-squared statistic that may be interpreted in the usual way

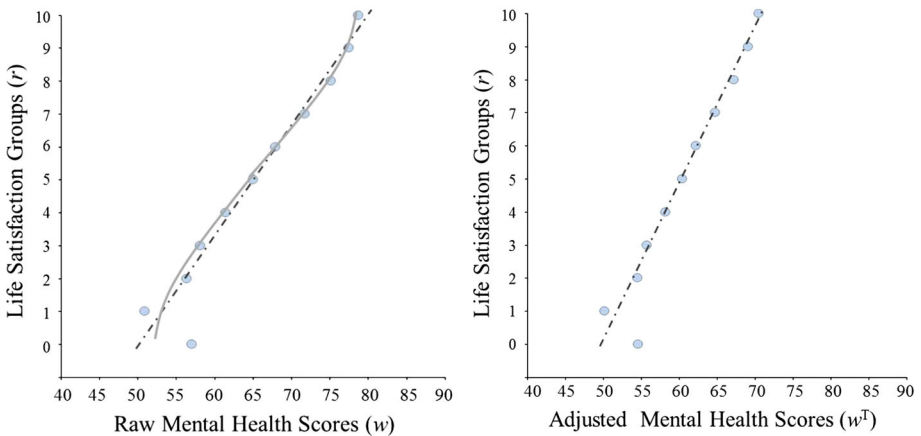


Fig. 3 Mental health characteristics of life satisfaction groups: within individuals

this assumption. Cardinal comparability is a stronger assumption, yet it is increasingly common, and justified almost exclusively on the basis of statistical convenience and with very little consideration about potential misuse of data. The argument for why cardinality

is desirable is clear, though potentially it weakens as information availability and statistical methods are improved. However, no identified work has looked directly and empirically at the question of why it might be reasonable, and this represents the key contribution of this paper.

The analysis presented here evaluates the ordinal and cardinal comparability of the eleven-point numeric life satisfaction scale, using the MH5 mental health index, treating the association between these two measures as the combination of two separate response functions which both reflect true wellbeing or utility. Consequently, we may observe something about the response function for life satisfaction via this association, with a level of specificity determined by what we are willing to assume about the response function for mental health. Because the mental health index is constructed and validated to fit the Rasch model, measurement theory decrees that the required adjustment of raw scores will produce a cardinal measure.

If one is prepared to accept a linear response function for adjusted mental health scores, the response function for life satisfaction is observable indirectly via the association between life satisfaction and adjusted mental health scores. If one is prepared only to accept that the response function for mental health is logistic in shape, but not fully known, then the shape of the association between life satisfaction and mental health scores provides some boundaries for what shape the response function for life satisfaction may take. Alternatively, if one is not prepared to make any assumptions whatsoever about the possible form of the response function for mental health, this association is unable to tell us anything useful about the response function for life satisfaction, and though this view cannot be dismissed it is clearly not advocated here.

First and foremost, the results presented here show that life satisfaction scores are ordinally distinct in terms of mental health, whether raw or adjusted ('cardinalised') data are used. This holds both across individuals and within individuals. The association between life satisfaction and mental health is here found to be quite strong, compared a range of other known correlates of subjective wellbeing, and approximately linear. Distances in raw mental health scores diminish at the edges of the life satisfaction scale, though this 'raw score bias' is reduced when adjusted scores are used. Visually, the association between life satisfaction and adjusted mental health scores appear very close to linear, both across and within individuals. However, equidistance of score-points is rejected by the statistical tests applied here, also when adjusted scores are used.

These results may be interpreted in a range of ways, as indicated. The only true evidence provided here pertains to the association between life satisfaction scores and raw or adjusted mental health scores. A number of factors determine the extent to which we can use this information to draw useful inferences about the metrics of the life satisfaction scale, and its underlying response function. One might argue that the degree of commonality between life satisfaction and mental health is too weak to be used as a basis for evaluating ordinal and cardinal comparability. A counter-argument is that these data are persistently noisy and that this degree of commonality is comparatively strong. Clearly, there is enough commonality for the hypothesis of ordered distinctness to hold.

Conservative data users may maintain that utility is and remains inherently unobservable and that cardinal comparability therefore cannot be justified. However, this view arguably also brings ordinal comparability into question, and thereby any useful evaluation of subjective wellbeing data, at least in economic analysis. A counter-argument could be that observation of people's subjective wellbeing and behaviour is consistent with ordinal comparability, because low satisfaction in certain spheres of life (like work and marriage) are generally associated with a higher probability of subsequent change (like change in

employment and separation or divorce).²¹ That is, we might accept ordinal comparability because the data behave in a way that is consistent with this assumption. If so, we ought to be equally willing to accept cardinal comparability when faced with evidence that the behaviour of subjective wellbeing data is consistent with what we expect of a cardinal measure.

There is an ongoing debate on metrics and inference in the broader literature, where conservatives warn against misrepresenting data which are not truly cardinal and producing invalid statistics (e.g. Katzner 1998) and others warn equally strongly against being too conservative and not making the best use of the information available (e.g. Guttman 1977). Guttman argues that researchers should select data analysis based on loss minimisation rather than on 'permission' in order to avoid wasteful and inefficient use of data. Otherwise, any metric which is not a purely objective measure, but is often treated as cardinal, must be demoted to ordinal status, including IQ, student grades, many health metrics, and a range of other index-like measures.

Consequently, scholars sympathetic to Guttman (1977) might consider the behaviour of these life satisfaction scores sufficiently consistent with the conditions of uniqueness and equidistance to provide a justifiable basis for both ordinal and cardinal comparability, both across and within individuals. Scholars more sympathetic to Oswald (2008) might maintain that utility is and remains unobservable, and that in spite of the sophisticated tools of psychometrics the assumption of cardinal comparability remains too heroic. The main purpose of this paper is not to argue for either side of this debate, but rather to provide much-needed information to enable data users to make more informed choices with respect to how they treat subjective wellbeing data, and in particular the eleven-point numeric life satisfaction scale. Nevertheless, all things considered, the evidence in favour of ordinal and cardinal comparability is considered compelling.

Acknowledgments This paper is dedicated to the memory of Paul W. Miller, who encouraged me to pursue this work on the metrics of subjective wellbeing. Several other scholars have provided valuable comments and guidance in the various developmental stages of this paper, including Robert Cummings, Stephen Pudney, Juerg Weber, Peter Robertson, David Butler, Paul Gerrans and David Andrich. I also wish to thank three anonymous referees for considered and valuable comments and suggestions. The study uses unit record data from the Household, Income and Labour Dynamics in Australia (HILDA) survey. The HILDA project was initiated and funded by the Australian Government Department of Families, Housing, Community Services and Indigenous Affairs (FaHCSIA) and is managed by the Melbourne Institute of Applied Economic and Social Research (MIAESR). The findings and views reported in this paper, as well as any mistakes or errors, are those of the author, and should not be attributed to any of the scholars listed in these acknowledgements or to FaHCSIA or MIAESR.

References

- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561–573.
- Bernoulli, D. (1738 [1954]). Exposition of a new theory on the measurement of risk. *Econometrica*, 22, 23–36.
- Blanchflower, D. G., & Oswald, A. J. (2004). Well-being over time in Britain and the USA. *Journal of Public Economics*, 88, 1359–1386.

²¹ Gardner and Oswald (2006) present evidence on marital satisfaction and subsequent likelihood of divorce. Clark et al. (2008) provide a list of studies comparing job satisfaction and job quits. Clark (2003) finds that those which are more negatively affected by becoming unemployed are also faster at finding new work.

- Blanchflower, D. G., & Oswald, A. J. (2005). *Happiness and the human development index: The paradox of Australia*. NBER Working Paper Series (No. 11416).
- Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist*, *61*(1), 27–41.
- Boyce, C. J. (2010). Understanding fixed effects in human well-being. *Journal of Economic Psychology*, *31*, 1–16.
- Boyce, C. J., & Wood, A. J. (2011). Personality and the marginal utility of income: Personality interacts with increases in household income to determine life satisfaction. *Journal of Economic Behavior & Organization*, *78*, 183–191.
- Brogden, H. E. (1977). The Rasch model, the law of comparative judgement and additive conjoint measurement. *Psychometrika*, *42*, 631–634.
- Clark, A. E. (2003). Unemployment as a social norm: Psychological evidence from panel data. *Journal of Labour Economics*, *21*(2), 323–351.
- Clark, A. E., Frijters, P., & Shields, M. A. (2008). Relative income, happiness, and utility: An explanation for the Easterlin Paradox and other puzzles. *Journal of Economic Literature*, *46*(1), 95–144.
- Crooker, K. J., & Near, J. P. (1998). Happiness and satisfaction: Measures of affect and cognition? *Social Indicators Research*, *44*, 195–224.
- Diener, E., & Lucas, R. E. (1999). Personality and subjective well-being. In D. Kahneman, E. Diener, & N. Schwartz (Eds.), *Well-being: The foundations of hedonic psychology* (pp. 213–229). New York: Sage.
- Edgeworth, Y. F. (1881 [1961]). *Mathematical psychics: An essay on the application of mathematics to the moral sciences*. New York: Augustus M. Kelly.
- Ferrer-i-Carbonell, A., & Frijters, P. (2004). How important is methodology for the estimates of the determinants of happiness? *The Economic Journal*, *114*(July), 641–659.
- Gardner, J., & Oswald, A. J. (2001). *Does money buy happiness? A longitudinal study using data on windfalls*. Warwick: Warwick University.
- Gardner, J., & Oswald, A. (2006). Do divorcing couples become happier by braking up? *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *169*(2), 319–336.
- Guttman, L. (1977). What is not what in statistics. *The Statistician*, *26*, 81–107.
- Headey, B., & Wooden, M. (2004). The effects of wealth and income on subjective well-being and ill-being. *Economic Record*, *80*(Special Issue), S24–S33.
- Hirschauer, N., Lehberger, M., & Musshoff, O. (2014). Happiness and utility in economic thought—or: What can we learn from happiness research for public policy analysis and public policy making? *Social Indicators Research*, *121*, 647–674.
- Katzner, D. W. (1998). The misuse of measurement in economics. *Metroeconomica*, *49*(1), 1–22.
- Kristoffersen, I. (2010). The metrics of subjective wellbeing: Cardinality, neutrality and additivity. *The Economic Record*, *86*(272), 98–123.
- Larsen, R. J., & Fredrickson, B. L. (1999). Measurement issues in emotional research. In D. Kahneman, E. Diener, & N. Schwarz (Eds.), *Well-being: The foundations of hedonic psychology*. New York: Sage.
- Lau, A. L. D. (2007). Measurement of subjective wellbeing: Cultural issues. In *9th Quality of Life Conference*. Deakin University, Melbourne.
- Layard, R., Mayraz, G., & Nickell, S. (2008). The marginal utility of income. *Journal of Public Economics*, *92*, 1846–1857.
- Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, *1*, 1–27.
- Masin, S. C., Zudini, V., & Antonelli, M. (2009). Early alternative derivations of Fechner's law. *Journal of the History of the Behavioral Sciences*, *45*(1), 56–65.
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, *63*(2), 81–97.
- Ng, Y.-K. (1996). Happiness surveys: Some comparability issues and an exploratory survey based on just perceivable increments. *Social Indicators Research*, *38*, 1–27.
- Ng, Y.-K. (2008). Happiness studies: Ways to improve comparability and some public policy implications. *The Economic Record*, *84*(265), 253–266.
- Oswald, A. (2008). On the curvature of the reporting function from objective reality to subjective feelings. *Economics Letters*, *100*(3), 369–372.
- Parducci, A. (1995). *Happiness, pleasure, and judgment: The contextual theory and its applications*. Hillsdale, NJ: Erlbaum.
- Perneger, T. V., & Bovier, P. A. (2001). *Application of the Rasch model to the SF36 mental health 5 item scale (MH5)*. ISPOR Sixth Annual International Meeting, Value In Health.
- Raczek, A. E., Ware, J. E., Bjorner, J. B., Gandek, B., Haley, S. M., Aaronson, N. K., et al. (1998). Comparisons of Rasch and summated rating scales constructed from SF-36 physical functioning items

- in seven countries: Results from the IQOLA Project. International Quality of Life Assessment. *Journal of Clinical Epidemiology*, 51(11), 1203–1214.
- Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* Berkeley, California, University of California Press.
- Sandvik, E., Diener, E., & Seidlitz, L. (1993). Subjective well-being: The convergence and stability of self-report and non-self-report measures. *Journal of Personality*, 61, 317–342.
- Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley.
- Schwartz, N. (1995). What respondents learn from questionnaires: The survey interview and the logic of conversation. *International Statistical Review*, 63, 153–177.
- Van Praag, B. M. S. (1991). Ordinal and cardinal utility: An integration of the two dimensions of the welfare concept. *Journal of Econometrics*, 50, 69–89.
- Van Praag, B. M. S., & Ferrier-i-Carbonell, A. (2004). *Happiness quantified*. New York: Oxford University Press.
- Ware, J. E., Snow, K. K., Kosinski, M., & Gandek, B. (2000). *SF-36 health survey: Manual and interpretation guide*. Lincoln, RI: QualityMetric Inc.
- Wright, B. D. (1997). Measurement for social science and education: History of social science measurement. <http://www.rasch.org/memo62.htm>.