Can Revealed-Preference Tradeoffs Be Inferred From Happiness Data? Evidence from Residency Choices

**** PRELIMINARY – PLEASE DO NOT QUOTE OR CIRCULATE ****

Daniel J. Benjamin
Cornell University and NBER

Ori Heffetz
Cornell University

Miles S. Kimball
University of Michigan and NBER

Alex Rees-Jones
Cornell University

First Draft: December 28, 2012
This Draft: January 2, 2013

Abstract

To what extent do marginal rates of substitution estimated from subjective well-being (SWB) data reflect the tradeoffs that individuals would deliberately choose to make? We survey 561 students from U.S. medical schools shortly after they submit their choice rankings over residencies to the National Resident Matching Program. Eliciting both choice rankings and anticipated-SWB rankings over residencies (using three common SWB measures), we find substantial differences between them in the implied tradeoffs between different features of the residencies. For example, while residency prestige and status weigh more in choice, expecting life to seem worthwhile during the residency weighs more in all SWB measures. Evaluative measures (life satisfaction and Cantril's ladder) imply tradeoffs closer to choice than does affective happiness. We further investigate the extent to which a multi-period happiness index and a multi-measure SWB index imply tradeoffs closer to choice. Despite the differences in implied tradeoffs, SWB questions predict pairwise choice reasonably well in our data, and often substantially better than alternative questions. We discuss implications of our findings for the use of SWB data in applied work.

JEL Classification: C81, D03, D69

Keywords: happiness, life satisfaction, subjective well-being, preference, utility, revealed preference

* We thank Al Roth for valuable early discussions and suggestions, and Aaron Bodoh-Creed, Sean Nicholson, Ted O'Donoghue, and Richard Thaler for valuable comments. We are grateful to participants at the Cornell Behavioral Economics Research Group, Cornell Behavioral/Experimental Lab Meetings, UCLA/UCSB Conference on Field Experiments, Michigan Retirement Research Center Annual Meeting, and Stanford Institute for Theoretical Economics, as well as seminar audience at Chicago Booth for helpful comments. We thank Allison Ettinger, Matt Hoffman, and Andrew Sung for outstanding research assistance. We are grateful to NIH/NIA grants R01-AG040787 and R01-AG020717-07 to the University of Michigan and T32-AG00186 to the NBER, and to the S. C. Johnson Graduate School of Management, for financial support.
E-mail: db468@cornell.edu, oh33@cornell.edu, mkimball@umich.edu, arr34@cornell.edu.

Since the marginal rate of substitution (MRS)—the rate at which substitution of one argument of the utility function for another would leave an individual indifferent—is a key quantity in many economic analyses, economists routinely attempt to estimate it. Traditionally, MRSs are estimated from choice data. Economists must resort to alternatives, however, in settings where the relevant choices are not observed (as is often the case when externalities, non-market goods, and government policies are involved) or where individuals' choices are likely to reflect mistakes. One increasingly used alternative source of data is subjective well-being (SWB) survey responses—most commonly, to questions about respondents' happiness, life satisfaction, or life's ranking on a ladder. In a typical application, a SWB measure is regressed on the quantities of a bundle of goods, and the tradeoff between a pair of goods is estimated as the ratio between their coefficients.[1] Under the assumption that the SWB measure proxies for utility—i.e., that the SWB measure is what individuals seek to maximize—the estimated tradeoff can be interpreted as the MRS between the two goods.

The purpose of this paper is to empirically explore the extent to which tradeoffs estimated from SWB data generate MRS estimates that reliably reflect individuals' preferences.[2] To that end, we elicit: (*a*) choice rankings over a set of options, in a setting where choice arguably reveals preferences; (*b*) the anticipated SWB consequences of the different choice options; and

[1] For example, in the public goods domain, Di Tella, MacCulloch and Oswald (2001) focus on a life satisfaction question to estimate the tradeoff between inflation and unemployment. In the externalities domain, a large literature on social comparisons uses a variety of SWB measures to estimate the MRS between own and others' income (for a recent review, see Clark, Frijters, and Shields, 2008). In the non-market goods domain, Deaton, Fortson, and Tortora (2010) use a variety of SWB measures, including the Cantril self-anchoring scale, to study the implied value of life in sub-Saharan Africa by comparing the coefficient on losing a family member with the coefficient on income. SWB data have been similarly used in a variety of settings, including to price noise (van Praag and Baarsma, 2005), informal care (van den Berg and Ferrer-i-Carbonell, 2007), the risk of floods (Luechinger and Raschky, 2009), air quality (Levinson, 2012), and the benefits of the Moving to Opportunity project (Ludwig et al., 2012), and to get a measure of tort compensation for the loss of family members (Oswald and Powdthavee, 2008).

[2] The literature seems to reflect a wide range of views regarding the relationship between SWB and preferences. On one extreme, Di Tella, MacCulloch and Oswald (2001) explicitly identify SWB measures with utility, and their estimates with iso-utility contours and MRSs: "The estimation describes preferences themselves." Perhaps on the other extreme, Deaton, Fortson, and Tortora (2010) discuss "well-being" rather than preferences, and explicitly consider the possibility that "the methods based on self-reported well-being do not tell us what we want to know." Moreover, they repeatedly point out that their ladder question implies dramatically different tradeoffs compared with their affective questions, and hence warn against using one SWB measure, or even a combination, as an exclusive guide. Committing to neither extreme, Frey and Stutzer (2002) hold in their *JEL* review: "Happiness is not identical to the traditional concept of utility in economics. It is, however, closely related… [it] is a valuable complementary approach… SWB can be considered a useful approximation to utility…"

(*c*) the expected quantities of the goods that comprise the relevant consumption bundle under each choice option. We estimate the tradeoffs between the goods implied by SWB and those implied by preferences, and we explore the relationship between them.

While the literature estimates the tradeoffs implied by *experienced* SWB, it is crucial for our purposes to compare choice tradeoffs with *anticipated* SWB tradeoffs in order to hold constant the conditions (including information and beliefs) under which choice is made. That way, we can attribute divergences to SWB not fully capturing the importance of certain goods in preferences. In contrast, divergences between choice and experienced SWB tradeoffs could result, for example, from mispredictions at the time of choice (e.g., Loewenstein, O'Donoghue, and Rabin, 2003; Gilbert, 2006).[3]

In section I we describe the setting of the choice we study: graduating U.S. medical students' preference rankings over residency programs. These preference rankings submitted by students to the National Resident Matching Program (NRMP), combined with the preference rankings over students submitted by the residency programs, determine which students are matched to which programs. This setting has a number of attractive features for our purposes: the matching mechanism is designed to be incentive-compatible; the choice is a deliberated, well-informed, and important career decision; the choice set is well-defined and straightforward to elicit; and due to a submission deadline, there is an identifiable moment in time when the decision is irreversibly made. We conduct a survey among a sample of 561 students from 23 U.S. medical schools shortly after they submit their residency preferences to the NRMP, so that our survey is conducted under the same conditions, including information set and beliefs, as the actual choice.

Section II describes our sample and survey design. We ask about each student's four most-preferred residency programs. In addition to eliciting each student's preference ranking over the four residencies as submitted to the NRMP, we also elicit her anticipated SWB rankings over the residencies, both during the residency and for the rest of her life. We focus on three

[3] It is logically possible that, despite the differences we find between anticipated-SWB tradeoffs and choice tradeoffs, experienced-SWB tradeoffs would nonetheless coincide with choice tradeoffs. This possibility seems very unlikely, however, since it requires that while individuals deliberately deviate, at the moment of making a choice, from what they believe would maximize their SWB, they somehow end up preferring maximized SWB anyway.

commonly-used SWB measures: happiness, life satisfaction, and a Cantril-ladder measure.[4] We also ask each student to rate each of the four residencies on each of nine features that we expected—based on our past research as well as on conversations with medical school officials and with past and present students—to be the most important determinants of program choice. These include the desirability of residency location, residency status and prestige, expected stress level, future career prospects, and future employment opportunities.

Section III reports our analyses and results. We model residencies as bundles of attributes, and use the choice- and SWB-rankings as alternative dependent variables in regressions where the independent variables are students' beliefs about these attributes. In our main analysis we compare the coefficients and coefficient ratios across regressions.

We find large and significant differences across the choice and SWB regressions in both the estimated marginal utilities and the implied tradeoffs. For example, relative to the choice-based estimates, the anticipated-SWB estimates underweight residency prestige and status, future career prospects, and desirability for the respondent's significant other, while overweighting social life, anxiety, and stress during the residency. We also find that our evaluative SWB measures—life satisfaction and ladder—generally yield results closer to the choice-based estimates than our more affective happiness measure. Our results are robust to plausible forms of measurement error and biases in survey response, and hold across empirical specifications and across subsets of our respondents.

We also explore whether multi-question SWB indices more accurately reflect revealed-preference tradeoffs. We consider three such indices: the first, a "3-SWB-measure" index, is a weighted sum of our three SWB questions; the second, a "4-interval-happiness" index, consists

[4] Examples of work using a variant of each of these three measures include: Di Tella, MacCulloch and Oswald's (2001) use of the Euro-barometer survey question: "On the whole, are you very satisfied, fairly satisfied, not very satisfied or not at all satisfied with the life you lead?"; Luttmer's (2005) use of the National Survey of Families and Households question: "Taking things all together, how would you say things are these days?" whose seven-point response scale ranges from "very unhappy" to "very happy"; and Deaton, Fortson, and Tortora's (2010) use of the Gallup World Poll question: "Please imagine a ladder with steps numbered from zero at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?"

Deaton et al.'s (2010) finding that different tradeoffs are implied by different SWB measures (see also Kahneman and Deaton, 2010) rules out the possibility that all SWB measures reliably reflect preferences at the same time. It leaves open the possibility that one of the measures, or a certain combination of measures, coincide with preferences; and it leaves open the question of whether certain measures come closer than others. We study these questions.

of happiness predictions for four time intervals that together cover the rest of a respondent's life; the third index combines the other two. While to the best of our knowledge no such indices have been used to date to estimate tradeoffs, we are motivated by the ideas, respectively, that well-being is multidimensional (e.g., Stiglitz, Sen, and Fitoussi, 2009; see Benjamin, Heffetz, Kimball, and Szembrot, 2012, for a first step toward implementation) and that well-being consists of instantaneous affect integrated over time (Kahneman, 2000). We estimate the optimal weights of the indices as best linear predictors of choice in our data; our indices are hence constructed to perform better than those likely to be constructed in realistic applications, where choice data are not available. We find that while some of these indices yield some tradeoffs that are closer to our choice-based MRSs than the indices' underlying questions, they do not in any meaningful way affect our main finding of substantial and highly statistically significant differences between SWB and choice tradeoffs.

In section IV, we explore an alternative application of SWB data, namely evaluating which of two concrete choice options makes an individual better off. Theoretical considerations suggest that, relative to estimating MRSs, researchers will often be on safer ground using SWB data for ordinal comparisons. We find that, indeed, despite the differences in implied tradeoffs between choice and SWB in our data, they often coincide in ordinal rankings. Excluding cases of anticipated-SWB indifference, anticipated happiness during the residency correctly predicts the binary choice ordering of programs 71% of the time; life satisfaction 77%; and our ladder question 80%.

We conclude in section V.

Our work builds upon and differs from past attempts to study the relationship between choice and SWB measures (Tversky and Griffin, 1991; Hsee, 1999; Hsee, Zhang, Yu, and Xi, 2003; Benjamin, Heffetz, Kimball, and Rees-Jones, 2012) in several important ways. First, while other studies document cases where choices do not maximize anticipated happiness, we are the first to study the implications for estimating MRSs.[5] Second, while other studies consider only single SWB questions, we also consider indices. Third, and most importantly, while existing work studies very small stakes or hypothetical choices, we present evidence on real choices in a

[5] With motivation similar to our own, Dolan and Metcalfe (2008) study the differences in estimated willingness to pay for an urban regeneration project as estimated using contingent valuation, SWB, and hedonic pricing methods. However, they observe experienced rather than anticipated SWB, and they do not directly observe ordinal preferences, instead inferring them from housing prices.

high-stakes field environment. Consequently, while in prior work it is arguable whether preferences are better reflected by choice or by anticipated SWB, in our setting there is a strong case for using choice as the gold standard measure of preferences.

I. Choice Setting: The National Resident Matching Program (NRMP)

I.A. Background

After graduating from a U.S. medical school, most students enroll in a residency program. The residency is a three- to seven-year period of training in a specialty such as anesthesiology, emergency medicine, family medicine, general surgery, internal medicine, pediatrics, and psychiatry. Students apply to programs at the beginning of their fourth year. In late fall programs invite selected students to visit and be interviewed. Students subsequently submit to the National Resident Matching Program (NRMP) their preferences over the programs where they interviewed, while programs submit their preferences over students. The NRMP determines the final allocation of students to residencies. In 2012, students were allowed to submit their preference ordering through the NRMP website between January 15 and February 22, and the resulting match was announced on March 16; among students graduating from non-homeopathic U.S. medical schools, 16,875 submitted their preference, and 15,712 (93%) ended up getting matched (NRMP, 2012).

The matching algorithm, described in detail in Roth and Peranson (1997, 1999), was designed to incentivize truthful preference reporting from students and to generate stable matches.[6] It is based on the student-proposing deferred acceptance algorithm of Gale and Shapley (1962), which is guaranteed to produce a stable match, and where truthful reporting is a weakly dominant strategy for students (Dubbins and Freedman, 1981; Roth, 1982). The original simple algorithm, however, could not accommodate certain requirements of the medical matching market (such as the need for couples to try and match to the same city). The needed modifications complicate the strategic motivations and allow the possibility that no stable match exists, but simulations in Roth and Peranson (1999) suggest that effectively all students remain incentivized to truthfully reveal their preferences.

[6] In two-sided matching problems, stability means that no student-and-program pair can be found where both the student and the program prefer to be matched to one another over their current matches.

I.B. Key Features for Our Study

For this paper, medical residency choices are an especially useful context for the following reasons:

Choice versus preferences: This setup may be as close as one can get to a setting where choices reveal preferences.[7] Residency choice is arguably one of the most important career-related decisions a medical student makes, with short- and long-term consequences for career path, geographic location, friendships, and family. Because of its importance, students deliberate over their decision for months and have a great deal of information and advising available to assist them in becoming well informed. Their submitted ranking is not visible to peers or residency programs, and hence, relative to many other decisions, the scope for strategic or signaling concerns is reduced. Finally and crucially, students are incentivized by the matching mechanism to report their true preference ranking.

Identifiable moment of choice: Unlike many other important life decisions, the NRMP submission has an identifiable moment when the decision is irreversibly committed. By surveying students shortly after they submit their preference ranking to the NRMP (and before they learn the match outcome), we elicit their SWB predictions under essentially the same information set and beliefs as at the moment of making the choice.[8]

Identifiable choice set and ranking: Unlike other decisions where observable choice data consist of the one chosen option, often with uncertainty as to the exact choice-set from which this option was chosen, here choice data consist of a ranking over a set of residencies. Therefore, we can elicit anticipated SWB and residency features over that same set of options. (As an additional benefit, observing a choice ranking over four options confers more statistical power than only observing which option was chosen from a set.)

Intertemporal tradeoff: A residency is expected to be a period of hard work, long hours, and intensive training, the benefits of which will be realized once the student becomes a

[7] Strictly speaking, what we refer to as our choice data are reports on choices; we do not directly observe the actual preference ranking submitted by students to the NRMP and have to rely on our survey respondents' reports. However, these reports seem very reliable. Among the 131 respondents who completed both our original and repeat surveys (see section II below), only 2 (1.5%) reported conflicting choice data. (Of the remaining 129 respondents, 5 had cross-survey differences in missing choice data but no conflicts; 2 seemed to have made easily-correctible data-entry mistakes in either survey; and 122 reported the exact same choices across the two surveys.)

[8] As discussed in section II.B below, figure 1 reports the distribution, over time, of NRMP submissions and survey responses. The median time between NRMP submission and survey response was 11 days.

practicing doctor. The investment aspect of the decision allows us to explore whether our affective SWB question—anticipated happiness during the residency—is better thought of as a measure of anticipated *instantaneous* well-being (a potential input into instantaneous, flow utility, the present discounted value of which determines choice) or as a measure of anticipated *discounted* well-being (resembling the assumption made in most applications using a happiness question). We consider and discuss this distinction in section III.C.

Preference heterogeneity: Residency choice offers rich variation in individuals' rankings of the same programs and in their assessments of programs' attributes; students' assessments of fit, locational preferences, and forecasts of the desirability of different programs are all reasonably idiosyncratic and are difficult to predict from objective, external data.[9,10] This also has advantages in terms of statistical power.

We also highlight a feature of residency choice that, while not advantageous for our purposes, is relevant in assessing our results and their external validity:

No monetary tradeoff: Residency choice is not well suited for studying tradeoffs with money. That is a disadvantage of our context, since using income as a numeraire is prominent in the literature. Our original intention was to use expected income under each residency for pricing the other residency attributes. However, in the process of designing the survey we learned—by being explicitly told by representatives of medical schools and by medical students we consulted—that expected income is largely unrelated to this decision. The primary determinant of expected income for medical students is their choice of *specialty*, a decision typically made years before choosing a residency. Indeed, most NRMP participants apply to residencies for a single specialty and hence should not expect their future income to vary meaningfully across their top choices. While pricing residency attributes in dollars would have been convenient, it is

[9] For some of our unsuccessful attempts to forecast residency choices in our data with objective, external measures such as characteristics of the city of residency, see web appendix table A8.

[10] Intuitions based on the academic job market for economists might suggest that students' preferences over residencies are largely determined by residencies' prestige rankings. While perceived prestige is strongly associated with preference rankings in our data, the relationship is far from perfect, and non-prestige factors are associated with preference even more strongly. This remains true when our survey measure for residency prestige and status is replaced by objective measures of prestige, such as data from the U.S. News and World Report Hospital Honor Roll. Conversations with medical students confirmed that prestige-independent issues such as locational preferences, spousal concerns, and assessment of "fit" play a large part in these decisions.

by no means crucial for our purposes; we instead focus on comparing MRSs and tradeoffs between the attributes directly. We elicited expected income in our survey anyway but do not analyze it in this paper.

II. Sample and Survey Design

II.A. Sample

From September 2011 to January 2012, we contacted all 122 fully-accredited U.S. medical schools by sending an email to a school representative (typically an Associate Dean of Student Affairs) and asking for permission to survey graduating medical students. We followed up with phone calls, further emails, and/or face-to-face meetings at the Association of American Medical Colleges Annual Meeting. As a result, 23 schools (19% of our initial list) agreed to participate in our study.[13] These 23 represent a wide range of class sizes (from 60 to 299 students in 2011) and locations, and they graduated a total of 3,224 students in 2011. Our survey appendix reproduces the initial email sent to schools, lists the participating schools, their class sizes, and the numbers of their students starting vs. completing our survey.

Between February 22 at 9pm EST (the deadline for submitting residency preferences) and March 3, students in participating schools received an email from their school's dean, student council representative, or registrar, inviting them to respond to our web survey by clicking on a link. The email is reproduced in the survey appendix. It explained, among other things, that "…The results of this study will provide better information on how medical students select residency programs, and can assist in the advising and preparation of future generations of students"; that the survey is estimated to take 15 minutes to complete; and that we offer participants at least a 1/50 chance to win an iPod.[14] Reminder emails were sent near the March 3 deadline. When the survey closed, at 11:59pm EST that day, we had received 579 complete responses (approximately 18% of the roughly 3,224 students contacted).[15] Our analysis is based

[13] A common reason schools gave us for not participating was that their students are already asked to participate in "too many" surveys.

[14] At the end of the survey, participants were thanked for their participation; were reminded that they have a 1/50 chance to win an iPod; and were asked to encourage their classmates to also participate. As an incentive for the latter, they were informed that we would increase the individual chance to win an iPod to 3/50 in schools with response rate of 70% or higher.

[15] In addition to the 579 complete responses, our survey had another 680 visits that did not result in a complete response. Of these, 284 (42%) exited before proceeding beyond the first page.

on the 561 who entered name and specialty information for at least two programs. Of those, 540 entered information for all four programs.

Participants who agreed to be re-contacted (when asked at the end of the survey) received, on a randomly-drawn date between March 7 and 9, another email inviting them to participate in a repeat survey, with a March 11 deadline. The repeat survey consisted of the same questions as the original survey, with a few new questions added. Comparing responses across these two waves allows us to assess the reliability of our measures, as we do below. 133 respondents completed the repeat survey; 131 of them provided information for at least two programs (23% of our main sample). The median time between completion of the original and the repeat surveys was 13 days.

II.B. Survey Design

Our survey appendix provides screenshots of our survey. Here we briefly summarize the important points. Following an introductory screen, respondents are asked: "Please enter the top four programs from the preference ordering you submitted to the NRMP."[16] Respondents separately enter program (e.g., "Massachusetts General Hospital") and specialty (e.g., "Anesthesiology").

Respondents are then asked: "On what date did you submit your rank order list to the NRMP?" Figure 1 reports the distributions of submission dates (lighter bars) and survey response dates (gray bars) among our 561 main-sample respondents. The median number of days between choice submission and response to our survey is 11. The figure also shows the subsequent distribution of response dates for the 131 main-sample respondents who participated in our repeat survey (darker bars).

On the next screen, respondents are asked about their relationship status and whether they are registered with the NRMP for a "dual match."[17] Their answer to the relationship question

[16] While the top four is not the entire preference ordering, it is likely to be the relevant portion of the list for our respondents. In 2012, 83.6% of NRMP participants graduating from U.S. medical schools were matched to one of their top four choices. (First choice: 54.1%; second: 14.9%; third: 9.1%; fourth: 5.5%; NRMP, 2012).

[17] The dual match is an option for couples trying to match to residencies simultaneously. The two submit a single list ranking pairs of programs. While 64% of our respondents indicate that they are either married or in a long-term relationship, only 7% are dual-match participants. As discussed in section III.B, our main results are robust to excluding them.

determines whether the question "On a scale from 1 to 100, how desirable is this residency for your spouse or significant other?" will be included as a residency attribute on a later screen.

Next, the following instructions appear on the screen:

> For the following section, you will be asked to individually consider the top four programs you ranked. For each of these possibilities, you will be asked to report your predictions on how attending that residency program will affect a variety of aspects of your life. Please answer as carefully and truthfully as possible.
> For some questions you will be asked to rate aspects on a 1-100 scale. Let 100 represent the absolute best possible outcome, 1 represent the absolute worst possible outcome, and 50 represent the midpoint.

The ranked residencies are then looped through in random order, and two screens appear for each residency. The first screen elicits respondents' rating of the residency, using the 1–100 scale, on the main three anticipated-SWB questions and on the nine residency attributes. The second screen includes questions about expected income that we do not use in this paper.

Table 1 reproduces the three anticipated-SWB questions and the nine attribute questions as they appear on the first screen below the instruction: "Thinking about how your life would be if you matriculate into the residency program in <specialty> at <program>, please answer the questions below." The SWB and residency questions are carefully designed to resemble each other as much as possible in terms of language and structure, and they appear on the screen mixed together as twelve questions in random order. The similar structure and symmetric treatment of the twelve questions on the screen allow us to compare the questions (in section IV below), without confounds due to question language or order, on how useful each one is as a single predictor of choice.

Mixed together and ordered here roughly by the time interval they refer to, the twelve SWB and attribute questions include: three affective measures that refer to *a typical day* during the residency (in the table these are labeled happiness, anxiety, and stress during residency); three evaluative/eudaimonic measures that refer more generally to the time during the residency (life satisfaction, social life, and worthwhile life during residency); one measure that refers implicitly to the time during the residency (desirability of location); one measure that refers implicitly to the time *after* the residency (future career prospects); one measure that simply refers to one's "life" (ladder); and three measures that come with no specification of period (residency prestige and status, control over life, and desirable for significant other).

Next, the top *three* residencies are cycled through again, in a new random order. For each residency we elicit anticipated happiness at different future time intervals (we provide more details when analyzing the resulting data, in section III.C below).

The survey concludes with a sequence of screens that include four questions regarding the relationship between a respondent's submitted NRMP ranking and her or his "true" preferences; a question regarding experiences with school representatives' attempts at manipulating the match; and questions about gender, age, college GPA, MCAT score, and Medical Licensing Examination scores. We explore these data in section III.B below. On the last screen, respondents are thanked for their participation and asked for permission to be contacted for the follow-up survey (428 agreed).

Figure 2 presents kernel density estimates of the distribution of our primary variables by residency rank, and appendix table A1 presents relevant summary statistics. As is visually clear, all have substantial variation, and many have clear differences in distribution across program ranks. For example, focusing attention on the three primary SWB measures (top row), it is clear that higher-ranked programs have higher mean SWB. Appendix table A2 presents the test-retest correlations of these variables, as calculated with the repeat survey. We view the high correlations of responses across waves as evidence that our survey measures elicit meaningful information.

III. Main Results

III.A. Single SWB Measures

As a first step in constructing choice-based and SWB-based MRS estimates, we estimate the marginal utilities implied by choice and by anticipated SWB.[18] The first four columns of table 2 report four separate regressions of, respectively, choice, anticipated happiness, anticipated life satisfaction, and anticipated ladder questions on the nine residency attributes. Each column estimates a rank-ordered logit model (Beggs, Cardell, and Hausman, 1981). This

[18] Of course, our marginal utility estimates and hence our MRS estimates may be biased, for example due to omitted variables. Since such biases would equally affect the choice-based and SWB-based estimates, our discussion is focused less on the point estimates themselves and more on whether they differ across choice and SWB measures.

model generalizes the standard binary-choice logit model to more than two ranked options.[19] The regressors are constructed by dividing the attribute variables by 100 (so the regressors range from 0.01 to 1). The coefficients can be interpreted analogously to standard logit coefficients: for any pair of residencies *A* and *B*, ceteris paribus, a one-unit increase in the difference in regressor j, $X_{i,A,j} - X_{i,B,j}$, is associated with a β_j increase in the log odds ratio of choosing *A* over *B*. We report a within-subject modification of McKelvey and Zavoina's R^2, measuring the fraction of within-subject variation of the latent index explained by the fitted model.[20]

Consider the two leftmost columns ("Choice" and "Happiness during residency"). The first row indicates that the coefficient on residency prestige and status is 2.5 in the choice regression and 0.0 in the happiness regression. This difference is strongly statistically significant (Wald test *p*-value = 0.000). To interpret these coefficients, consider their implication for the ranking of two residency programs that are identical in all measured dimensions except for a 10-point difference in their prestige and status on the survey's 100-point scale. The choice coefficient implies that the probability of choosing the more prestigious program would be $\frac{e^{2.5*10/100}}{e^{2.5*10/100}+1}$ = 56%. The happiness coefficient implies that the probability of ranking the more prestigious program higher on anticipated happiness would be 50%.

Our estimate of the marginal utility of the prestige and status attributed to a residency hence strongly depends on whether it is estimated from choice or from anticipated happiness. Examining the rest of the coefficient pairs across the choice and happiness columns reveals more such differences. With the exception of control over life, they are all statistically significant at the 10% level. Five of the differences are significant at the 1% level: in addition to residency prestige and status, also desirability of location, future career prospects, and desirability for significant other are associated significantly more with choice than with anticipated-happiness,

---

[19] Rank-ordered logit assumes that individual *i*'s ordinal ranking of residencies $r \in \{1, \ldots R\}$ is rationalized by a random latent index, $U_{ir} = \beta_{\boldsymbol{X}}\mathbf{X}_{ir} + \varepsilon_{ir}$. The unobserved error term is assumed to follow a type I extreme value distribution, yielding a closed-form solution to the implied maximum-likelihood problem.

[20] We modify the R^2 measure of McKelvey and Zavoina (1975) by demeaning the predicted index value $\widehat{U}_{ir}$ at the person level:

$$\frac{\widehat{Var}(\widehat{U}_{ir} - \overline{U}_i)}{\widehat{Var}(\widehat{U}_{ir} - \overline{U}_i) + Var(\varepsilon_{ir})}.$$

This ratio captures the fraction of within-subject variation in (latent) utility coming from the estimated, deterministic component, giving a measure of fit intuitively similar to standard R^2.

while the reverse is true for social life during the residency. As reported in the table's bottom row, joint equality of coefficients between the two columns is strongly rejected.

Looking at the next two columns ("Life satisfaction during residency" and "Ladder") reveals that with few exceptions, these two measures' marginal utility estimates lie between those of choice and those of happiness. These two evaluative measures seem at times closer to happiness, an affective measure, and at times closer to choice: while on social life during the residency, for example, the two are indistinguishable from happiness, all with coefficients larger than that on choice, on desirability of location they are indistinguishable from choice, with coefficients much larger than that on happiness. Across the rows, all the ladder estimates appear closer to choice than the life satisfaction estimates; statistically, however, we cannot distinguish the two evaluative measures from each other. Indeed, Wald tests of the joint equality of coefficients between any pair among the four columns strongly reject the null of equality (p = 0.000), with the exception of the life satisfaction and ladder pair (p = 0.52).

To what extent do these differences in marginal utility estimates translate to differences in estimated tradeoffs? To answer this question, the coefficients of these regressions must be normalized in a way that allows their implied tradeoffs to be directly compared across columns. One possible such normalization is reported in table 3. The table presents the ratio of each coefficient from table 2 to the average absolute value of coefficients in its table 2 column. With this normalization, for example, a higher coefficient on an attribute in the choice column relative to the happiness column means that on average, the MRS between another attribute and this one will be lower in the choice column.

Comparing table 3's column 1 with columns 2–4 reveals dramatic differences between the tradeoffs implied by choice and those implied by the different anticipated SWB measures. For example, in the first row, residency prestige and status's marginal utility in the choice column is 1.4 times the average of the nine attributes' marginal utilities. However, with any of the three anticipated SWB measures, prestige and status's estimated marginal utilities are below average, ranging from 0.0 to 0.4 times the average.

Examining other attributes, we again see clear differences between choice and all SWB measures in a number of cases. Appendix table A3 reports cross-regression differences in coefficient ratios and the p-values of each difference. Relative to the choice-based estimates, all three SWB measures underweight residency prestige and status and desirability for significant

other, and overweight the importance of social life and life seeming worthwhile during the residency. There are also significant differences for the other attributes, but they appear to be less systematic. As reported in table 3's bottom row, we again easily reject joint equality of (normalized) coefficients between any of the three SWB measures and choice.[21] And as in table 2, the life satisfaction and ladder columns appear similar to each other, with all estimates in between those in the choice and those in the happiness columns. Considered jointly, the coefficients in both the life satisfaction and ladder columns are again statistically different from the happiness column ($p = 0.000$) but are not distinguishable from each other ($p = 0.63$).

III.B Robustness

Bias in survey response: A halo effect or cognitive dissonance could cause respondents to modify their subjective assessments of either anticipated SWB or residency attributes (or both) in order to rationalize the choice order they reported earlier in the survey. However, note that such a bias in the ratings of the residency attributes, while biasing upward the coefficients in the choice column, cannot in itself explain the *differences* in coefficients across columns. In fact, such a bias in the ratings of anticipated SWB measures would *increase* the concordance between the SWB-based rankings and the choice ranking, biasing downward any choice-SWB differences across the columns. Therefore, the differences we do observe should be viewed as a lower bound on the actual divergence between anticipated-SWB and choice rankings.

Measurement error: If anticipated SWB is a noisy measure of choice utility, then differences in coefficients across our regression columns are to be expected. However, such measurement error predicts that, when considering two bundles of residency attributes, the probabilities of either being higher ranked will always be closer to 50-50 when calculated from anticipated SWB. To test this implication, we consider each two-residency comparison in our data and calculate the predicted probabilities implied by the residency attributes in the estimated choice and SWB models. We find that relative to the estimated probabilities for the choice ordering, the estimated probabilities for the SWB ordering are closer to 50-50 only 36% of the time for happiness, 42% of the time for life satisfaction, and 49% of the time for ladder. This evidence is inconsistent with measurement error in SWB being the sole difference between

[21] We test for joint equality by nesting both dependent variables into a single regression and using the delta method to recover the normalized coefficients and the covariance matrix. We use these to conduct a multivariate Wald test of the joint equality of each normalized coefficient across columns.

measures. While some degree of measurement error in our SWB variables is surely present, it cannot generate our main results.

Econometric approach: The estimates in tables 2 and 3 are based on a rank-ordered logit model. The model is desirable for its comparability to choice-based methods, and it requires no assumption about similar use of the dependent-variable rating scales across respondents. It is different from the typical approach taken in the literature, where dependent-variable scale use is assumed to be identical across people (or the same up to differences in means as in fixed-effects regressions).[22] For comparability, we conduct OLS regressions with respondent fixed effects as well as ordered logit regressions, reported in appendix tables A4 and A5. These regressions yield estimates similar to the rank-ordered logit regressions, and the discussion in the previous subsection is robust to this specification.

Heterogeneity: The analysis above may be thought of as assuming a representative agent. Heterogeneity in marginal utilities in itself could not drive our primary results that SWB measures yield different tradeoff estimates compared with choice. However, it is possible that such results are driven by a particular subpopulation, and that for many or most in the sample, the tradeoffs represented by their anticipated SWB are similar to those implied by their choices. To assess this possibility, we cut the sample along various respondent characteristics. For each sample cut, we re-estimate table 2 and test if each SWB column remains statistically different from the choice column. We reject the null hypothesis of identical marginal utility estimates at the 5% level for all such cross-column comparisons when restricting the sample along the following dimensions: relationship status, gender, above and below median MCAT scores, above and below median age, whether or not the respondent believes the NRMP submission represented her true preferences (83% of our sample believe it did), agreed to be re-contacted for the follow-up survey (76%), completed the follow-up survey (23%), excluding dual-match participants (7%), and excluding those who report manipulation attempts by schools (3%).[23] These tests suggest that our main results are pervasive across subgroups within our sample.

[22] This approach invokes more assumptions and generally has more weaknesses; for example, monotone transformations of the outcome variable would change estimates without leading to different preferences or implied welfare orderings, and it must be assumed that all respondents use the scales in the same way.
[23] Given that the mechanism for the NRMP was designed with incentive compatibility in mind, it might be surprising that only 83% of our sample indicate they believed their submission represented their true preferences. Of the remainder, however, only 5% indicate that they chose their list strategically, and less than 1% indicate that they felt they made a mistake. The remaining 11% indicate another reason, and are

III.C. Multi-Question SWB Indices

Our results thus far suggest that none of our single-question anticipated-SWB measures coincide with choice utility. However, two distinct hypotheses separately imply that *combinations* of questions may better capture choice utility. We now explore these two hypotheses.

Happiness as flow utility: When a survey respondent reports feeling happy, is her report better viewed as reflecting an instantaneous flow of well-being or as reflecting her feeling, at the moment of reporting, about her expected present discounted value of such well-being flows?

The former view has certain intuitive appeal; however, it significantly complicates the use of happiness questions for estimating tradeoffs. With that view, the interpretation of the happiness regressions as estimating choice-utility MRSs would be defensible only in situations with no significant intertemporal dimension.

To explore whether anticipated happiness would better reflect choice if it integrated happiness predictions regarding the full expected horizon of life, rather than regarding only the residency years, we elicit such additional predictions in our survey. As mentioned in section II.B above, after responding to questions about each of the top four residencies, the respondents cycle again through the top three, in a new random order. They are instructed as follows:

> For the following section, you will again be asked to individually consider the top three programs you ranked. For each of these possibilities, you will be asked to report your predictions on how attending that residency program will affect your happiness during different periods of your life. Please answer as carefully and truthfully as possible.

For each residency, respondents see a screen with questions. Three primary questions read: "On a scale from 1 to 100, how happy do you think you would be on average [during the first ten years of your career]/[for the remainder of your career before retirement]/[after retirement]?" Each is followed by questions assessing the uncertainty of the forecast.

Aggregating such questions into a discounted happiness index requires weighting them by appropriate discount factors (taking into account the different lengths of their respective intervals). In a field setting where choice data are not available, the researcher would have to

---

free to explain in a free-response textbox. Most such explanations point to constraints based on family preferences or location, perhaps indicating that the preferences we estimate for these respondents are best understood as those of their household, as opposed to themselves as individuals.

choose weights based on beliefs regarding the discount factor. Since we have choice data, we instead conduct a rank-ordered logit regression predicting choice with our four anticipated happiness questions, and we use the estimated latent index coefficients as our weights. In our data, this is the best linear index that could be constructed for predicting choice and hence represents a best-case scenario—from a choice-prediction perspective—for a discounted happiness measure that might be used in a realistic application.

This regression for constructing the index is reported in column 1 of table 4. The coefficients on the happiness variables are roughly declining over time, in spite of the increase in time-interval length, consistent with discounting.[24] However, the McKelvey and Zavoina R^2 of 0.13 indicates relatively low goodness-of-fit, hinting that the use of the recovered index may still omit significant amounts of choice-relevant information.

Returning to tables 2 and 3, in column 5 we use this multi-period anticipated-happiness index as the dependent variable ("4-interval-happiness index").[25] Table 2 shows that while column 5 is closer than column 2 (happiness during residency) to column 1 (choice) on many—but not all—of the coefficients, its coefficients still show substantial differences from the marginal utility estimates in column 1 (joint significance of differences $p = 0.000$ between columns 1 and 5; $p = 0.13$ between columns 2 and 5). Table 3 translates the coefficients to tradeoffs and conveys the same general picture ($p = 0.000$ and $p = 0.08$, respectively, in these two joint significance tests). Moreover, in both tables, columns 3 and 4—life satisfaction and ladder—seem in general closer than column 5 to column 1 (both columns 3 and 4 are statistically different from column 5, with $p = 0.01$ or less in either table).

In summary, while we may find some support for the "happiness as flow well-being" hypothesis, our four-time-interval anticipated-happiness index is still far from yielding reliable

[24] While we do not know the exact length of three of the time intervals, we can calculate them roughly. The during-the-residency happiness measure would typically cover five years starting from the present. By definition, we know that the first-ten-years-of-career measure covers the ten years that follow. Since the average age in our sample is 27, the rest-of-career measure is expected to cover roughly another 23 years until retirement (= 65–27–5–10). With life expectancy roughly 80 years at that age, the after-retirement measure would cover on average another 15 years. Hence, relative to the during-the-residency measure, the first-ten-years-of-career is roughly twice as long, and the last two time windows are roughly three to five times as long.

[25] Since the three beyond-residency anticipated-happiness questions are elicited for only the top three residency choices, the estimates in column 5 in tables 2 and 3 are based on a subset of the data columns 1–4 are based on. In web appendix tables A9 and A10 we reproduce the two tables limiting them to the data column 5 is based on, and show that our results are not driven by this potential selection issue.

MRS estimates. In particular, the index does not seem to do better than our single-question evaluative SWB measures.

Multidimensional SWB: Although much of the economics literature treats different SWB questions as interchangeable, several recent papers find that different questions have different correlates and argue that they capture distinct components of well-being.[26] To the extent that well-being is multidimensional, a multi-question SWB index might yield tradeoff estimates that are closer to our choice-based MRS estimates than those yielded by any single measure.

To explore this possibility, we construct a "3-SWB-measure" index from our main three SWB questions, and a "6-SWB-question" index by additionally including the three beyond-residency happiness questions (from the 4-interval-happiness index above). To maximize the predictive power of the index for choice, we again use as weights the coefficients estimated in first-stage regressions of choice on the components of each index.

Columns 2 and 3 of table 4 report our first-stage regressions. In both regressions the coefficient on happiness during the residency is indistinguishable from zero, and is substantially smaller than the corresponding coefficient in column 1 as well as than the coefficients on the two evaluative measures in columns 2 and 3 (life satisfaction during the residency and ladder). This suggests that once the latter two measures are controlled for, happiness during the residency contributes significantly less to predicting choice. The fit of the indices in columns 2 and 3, as measured by the McKelvey and Zavoina R^2, is substantially better than in column 1, suggesting that the two multidimensional SWB indices might be closer to choice than the multi-period happiness index.

Returning to tables 2 and 3, columns 6 and 7 in the two tables use, respectively, each of the two SWB indices as the dependent variable in the regression. We easily reject, in both tables, joint equality of coefficients between each of the two multi-SWB regressions and: choice (see tables' bottom row), happiness ($p = 0.000$), and, less strongly, 4-interval-happiness index ($p =$ 0.06 or less). Nonetheless, we cannot distinguish them from each other or from either life satisfaction of ladder (p-values range from 0.15 to 0.97).[27]

[26] The view that different SWB measures are interchangeable seems pervasive. For example, in a very recent paper, Ludwig et al. (2012) state that happiness questions yield results similar to those from general life satisfaction questions, and that "both provide global retrospective assessments of how people think their lives are going."

[27] It may seem surprising that, relative to single-question life satisfaction or ladder questions, the two indices do not in general yield coefficients and tradeoff estimates that are closer to those based on choice,

To summarize, we interpret our findings as potentially supporting the "multidimensional well-being" hypothesis, but even so, indices that incorporate multiple SWB measures still do not recover choice-based MRS estimates or do better than our single-question evaluative SWB measures. Of course, the SWB measures we include in these indices are far from exhausting every conceivable measurable dimension (and timing) of well-being, and hence we cannot reject that some sufficiently rich set of SWB questions might provide a fully adequate utility index. Nonetheless, since the measures we use are modeled after those most common in existing social surveys and applied research, our results suggest that a simple extension of current practices—using a linear combination of commonly-used SWB measures—would not be a substantial improvement for estimating utility tradeoffs.

IV. From Slopes to Orderings: Predicting Choice Ranking from Anticipated-SWB Ranking

Our results suggest that there are substantial differences between the MRSs implied by widely-used SWB measures and those revealed by choices. While this finding calls into question the practice of using SWB data to estimate tradeoffs—for example, to price things by estimating their tradeoff with income—SWB data could instead be used less ambitiously for assessing which among a set of options is most preferred. For SWB thus used to reflect preferences, it is sufficient that SWB levels would imply the same ranking of the options as choices do.

Figure 3 illustrates this simple point for the case of two goods. The solid line represents an individual's iso-utility curve, while the dashed line represents her iso-SWB curve. The respective slopes, or MRSs, at choice option A differ: SWB tradeoffs do not reflect preference tradeoffs. Indeed, while option A is preferred to choice option C, SWB is higher in C than in A; a SWB-based pairwise comparison of the options would in that comparison favor the wrong option from a preference point of view. At the same time, option B is both preferred to and ranks higher SWB-wise than option A; if the pair of options under consideration is A versus B—or, indeed, A versus any of many other options in the region where B lies—SWB data would yield the right choice.

The figure illustrates that with different implied tradeoffs, whether SWB data still yield a ranking of choice options that coincides with preference ranking depends on the exact shapes of

since the indices are better predictors of choice by construction. This finding is directly related to the fact that while a measure may be highly correlated with choice, it may not necessarily yield tradeoff estimates similar to those implied by choice. See section IV for discussion.

the iso-utility and iso-SWB curves, as well as on the orientation of the considered options relative to each other. In the specific case illustrated in the figure, an increase in a good has a positively-signed effect on both utility and SWB—a seemingly reasonable assumption in many cases although far from guaranteed. In that case, when one option vector-dominates the other, SWB rankings and preference rankings would coincide. More generally, in the absence of complete information regarding both curve shapes and bundle composition, the usefulness of SWB data as predicting pairwise choice is at least in part an empirical question. In this section we explore this question in our data.

The specific measure of usefulness we employ is the answer to the following question: if we randomly draw a pair of programs from all possible single-respondent pairs (i.e., all pairs of options that consist of a single respondent's top-ranked programs entered in our survey, added over all of our respondents), how well could we predict a respondent's binary choice (i.e., which program in the pair was ranked higher) from that respondent's responses to each of the SWB or non-SWB questions? Table 5 presents our answer.

The top panel of the table shows that by this "best pairwise predictor" measure, the ladder question is the most useful among our three SWB and nine residency attribute questions. Its ranking correctly predicts pairwise choice ranking in 80% of the cases in which it is informative—i.e., of the cases in which it is not equal across the two options, which in turn correspond to 82% of all cases.[28] The next best predictor *for respondents in a relationship* is desirability to one's partner, which predicts choice correctly in 77% of the cases in which it is informative—84% of cases for respondents in a relationship. But remember that only 64% of our respondents are in such relationship. The second best predictor for those not in a relationship (and third best for those in a relationship) is the life satisfaction question: its ranking coincides with choice in 77% of the cases in which it is informative—these cases in turn reflect 77% of all

[28] To rank our survey questions by this notion of pairwise predictive power, we calculate the percentage of total cases in which using a question as a proxy for choice would yield a correct choice ranking, with the assumption of predicting a correct choice ranking on average half the time when the survey question is uninformative. Thus, for example, the ladder question predicts choice correctly 80% (column 4 in table 5) of the 82% (100% – column 2) of cases in which it is informative, or 65% (column 1) of all cases; and choice could on average be guessed correctly 9% of the additional 18% of cases that are uninformative, adding up to 74% correct prediction out of all cases. Equivalently, the ladder predicts the wrong choice in 17% (column 3) of all cases, and choice would be guessed incorrectly in another 9% of all choices, to an average total of 26% of cases. In contrast, a similar calculation suggests that anxiety would yield correct choice prediction in less than 53% of all cases—just slightly better than chance.

cases. The next best predictors are desirability of location (71% correct of the informative 86% of the cases), happiness during the residency (71% correct of 73% informative), life seeming worthwhile (73% of 60%), career prospects (70% of 70%), prestige (67% of 84%), and social life (65% correct of 80% informative). At the bottom, anxiety (53% of 71%) and stress (54% of 74%) during the residency, and control over life (57% of 70%) do only slightly better than a 50-50 guess.

The middle panel of table 5 adds to the twelve measures above the three beyond-residency happiness questions. As the time interval goes further into the future, the correct prediction rate is seen to decrease and, crucially, the percentage of uninformative cases (column 2) increases. The latter is very high, at 53%, even for the measure that is closest in time among the three—happiness in the first ten years of one's career—rendering these measures of limited usefulness as single-question predictors of pairwise choices.

Finally, for comparison with these single-question measures, the bottom panel of the table reports on the pairwise predictive success of our three multi-question indices (discussed in III.C) and two additional indices that combine the nine attribute questions into the multidimensional SWB indices.[29] The 4-interval-happiness index is seen informative more often than the happiness-during-the-residency question: uninformative cases drop from 27% to 10%. However, this comes at the cost of a drop in correct prediction rate among informative cases, from 71% to 69%. Using our measure of usefulness, hence, while this happiness index does slightly better than our single happiness question above, it is still not as useful as our desirability-of-location or more useful questions (including life satisfaction and ladder). The rest of the indices, which are based on increasing numbers of questions—3, 6, 12, and 15—show relatively high and increasing correct prediction rates—77%, 78%, 81%, and 82%—and low and decreasing percent of uninformative cases—3%, 2%, 0% and 0%.

These results may be informative to practitioners. They suggest that in terms of getting the direction of certain binary choices right, single-question evaluative anticipated-SWB measures such as ladder and life satisfaction, as well as a desirable-for-significant-other question when relevant, may be useful. Naturally, combining them with other questions into multi-question indices that are constructed to produce best-linear-predictor indices in a given data set

[29] The weights in these two additional indices are estimated in the same way other weights are estimated; the relevant regressions are reported in web appendix table A11.

improves their predictive usefulness—in our data, mainly by eliminating uninformative cases. However, we do not know whether the indices that can be constructed in specific applications and in the absence of choice data do as well; and whether the cost of eliciting additional questions to construct these indices is justified by the improvement in prediction.

In web appendix tables A6 and A7 we report two additional versions of table 5, restricting the underlying data to two respective subsets of pairwise program comparisons: only first- versus second-ranked programs, and only first- versus third-ranked programs. We find, as expected, that all of our measures are better predictors of choice in the latter than in the former (for example, ladder's correct prediction rate increases from 78% to 87%). This finding is consistent with another practical implication of the case illustrated in figure 3, namely that SWB may in general be more likely to favor the preferred option the farther from indifference the two considered options are.

V. Concluding Remarks

Economists have gained many important insights, and are likely to continue gaining new ones, by regressing SWB measures on bundles of goods and comparing the estimated coefficients. From the point of view of utility theory, however, the aspects of well-being captured by traditional SWB measures should in principle not be treated differently from the goods these measures are often regressed on. Indeed, utility theory views both groups as right hand side variables, i.e., as potential utility inputs and their correlates. Of course, traditional SWB measures may well represent *important* utility inputs—more important than many other goods. Such a view is consistent with the relatively high correlations in our data between anticipated-SWB measures and choice. But, as has been suggested by some researchers, it seems unlikely that one SWB question or even a combination of a small number of them would capture enough of the important inputs to be sufficient as an all-purpose utility proxy. In particular, our finding that in our data the tradeoffs implied by anticipated SWB differ in important ways from the MRSs implied by deliberated choice serves as a warning sign against making the working assumption that SWB = utility.

The evaluative measures we explore—life satisfaction and Cantril's ladder—yield tradeoff estimates that are significantly closer to those yielded by choice than our affective happiness measure, even when the latter attempts to integrate several time intervals. While

applied researchers may find such findings useful, one should remember that many other issues that we do not touch upon in this paper still bedevil the measurement of SWB (see, e.g., Adler, 2012).

Finally, our evidence is limited to one specific context, and the nine residency attributes that constitute our bundle of goods are far from exhaustive. Clearly, more evidence is needed before one can seek general conclusions regarding the magnitude or even the sign of the bias in specific MRSs when estimated from SWB rather than from choice data. That said, two of our findings in this paper seem closely aligned with previous findings from stated preference data (in very different contexts), and may hence be worth repeating. First, compared with affective happiness measures, life satisfaction measures are consistently closer to choice (in this paper) and to stated choice (in Benjamin, Heffetz, Kimball, and Rees-Jones, 2012; and in Benjamin, Heffetz, Kimball, and Szembrot, 2012). Second, all three papers find that measures of family well-being—family happiness (in the two papers cited above) and residency desirability to one's spouse or significant other (in the present paper)—are underweighted in SWB relative to their weight in choice or stated choice. Exploring these emerging generalized conclusions in new settings and with different empirical approaches would be a natural direction for future research.[30]

References

Adler, Matthew D. 2012. "Happiness Surveys and Public Policy: What's the Use?" University of Pennsylvania Law School Research Paper 12–36, http://ssrn.com/abstract=2076539.

Beggs, S., S. Cardell, and J. Hausman. 1981. "Assessing the Potential Demand for Electric Cars." *Journal of Econometrics,* 16: 1–19.

[30] At the same time, some of our findings in the present paper differ from past findings. For example, our finding that our Cantril ladder question is closer to choice than our happiness question and is as close to choice as our life satisfaction question, and our finding that status and prestige are strongly positively associated with choice may at first appear inconsistent with the findings in Benjamin, Heffetz, Kimball, and Szembrot (2012) that happiness and especially life satisfaction measures are significantly more correlated with stated choice than a Cantril ladder measure, and that social status is weakly negatively associated with stated choice. However, the different findings may be explained by the different methodology and purpose of that previous paper; see discussion in section IV.B there, where the authors conjecture that their findings "may reflect respondents' answering our stated preference question in terms of their meta-preferences or laundered preferences."

Benjamin, Daniel J., Ori Heffetz, Miles S. Kimball, and Alex Rees-Jones. 2012. "What Do You Think Would Make You Happier? What Do You Think You Would Choose?" *American Economic Review*, 102(5): 2083–2110.

Benjamin, Daniel J., Ori Heffetz, Miles S. Kimball, and Nichole Szembrot. 2012. "Beyond Happiness and Satisfaction: Toward Well-Being Indices Based on Stated Preference." NBER Working Paper No. 18374.

Clark, Andrew, Paul Fritters and Michael Shields. 2008. "Relative Income, Happiness and Utility: An Explanation for the Easterlin Paradox and Other Puzzles." *Journal of Economic Literature*, 46(1): 95–144.

Deaton, Angus, Jane Fortson, and Robert Tortora. 2010. "Life (Evaluation), HIV/AIDS, and Death in Africa." In *International Differences in Well-Being*, edited by Ed Diener, John Helliwell, and Daniel Kahneman, Oxford*:* Oxford University Press, 105–136.

Di Tella, Rafael, Robert J. MacCulloch, and Andrew J. Oswald. 2001. "Preferences over Inflation and Unemployment: Evidence from Surveys of Happiness." *American Economic Review,* 91(1): 335–341.

Dolan, Paul and Robert Metcalfe. 2008. "Comparing Willingness-To-Pay and Subjective Well-Being in the Context of Non-Market Goods." CEP Discussion Paper No 890.

Dubbins, Lester and David Freedman. 1981. "Machiavelli and the Gale-Shapley Algorithm." *The American Mathematical Monthly.* 88(7): 485–494.

Frey, Bruno and Alois Stutzer. 2002. "What Can Economists Learn from Happiness Research?" *Journal of Economic Literature,* 40(2): 402–435.

Gale, David and Lloyd Shapley. 1962. "College Admissions and the Stability of Marriage." *American Mathematical Monthly,* 69: 9–15.

Gilbert, Daniel. 2006. *Stumbling on Happiness*. New York: Knopf.

Hsee, Christopher K. 1999. "Value-Seeking and Prediction-Decision Inconsistency: Why Don't People Take What They Predict They'll Like the Most?" *Psychonomic Bulletin and Review*, 6(4): 555–561.

Hsee, Christopher K., Jiao Zhang, Fang Yu, and Yiheng Xi. 2003. "Lay Rationalism and Inconsistency Between Predicted Experience and Decision." *Journal of Behavioral Decision Making*, 16: 257–272.

Kahneman, Daniel. 2000. "Experienced Utility and Objective Happiness: A Moment-Based Approach." *Choices, Values, and Frames*, ed. Kahneman, Daniel, and Amos Tversky. Cambridge, UK: Cambridge University Press.

Kahneman, Daniel, and Angus S. Deaton. 2010. "High Income Improves Evaluation of Life but not Emotional Well-Being." *Proceedings of the National Academy of Sciences*, 107(38): 16489–16493.

Levinson, Arik. 2012. "Valuing Public Goods Using Happiness Data: The Case of Air Quality." *Journal of Public Economics*, 96: 869–880.

Loewenstein, George, Ted O'Donoghue, and Matthew Rabin. 2003. "Projection Bias in Predicting Future Utility." *Quarterly Journal of Economics* 118 (4): 1209–48.

Ludwig, Jens, Greg J. Duncan, Lisa A. Gennetian, Lawrence F. Katz, Ronald C. Kessler, Jeffrey R. Kling, Lisa Sonbonmatsu. 2012. "Neighborhood Effects on the Long-Term Well-Being of Low-Income Adults." *Science,* 337: 1505–1510.

Luechinger, Simon, and Paul A. Raschky. 2009. "Valuing Flood Disasters Using the Life Satisfaction Approach." *Journal of Public Economics*, 93: 620–633.

Luttmer, Erzo. 2005. "Neighbors as Negatives: Relative Earnings and Well-being." *Quarterly Journal of Economics,* 120(3): 963–1002.

McKelvey, Richard and William Zavoina. 1975. "A Statistical Model for the Analysis of Ordinal Level Dependent Variables." *Journal of Mathematical Sociology*, 4: 103–120.

National Resident Matching Program. 2012. "National Resident Matching Program, Results and Data: 2012 Main Residency Match℠." National Resident Matching Program, Washington, DC.

Oswald, Andrew, and Nattavudh Powdthavee. 2008. "Does Happiness Adapt? A Longitudinal Study of Disability with Implications for Economists and Judges." *Journal of Public Economics*, 92: 1061–1077.

Roth, Alvin E. 1982. "The Economics of Matching: Stability and Incentives." *Mathematics of Operations Research*, *7*: 617–28.

Roth, Alvin and Elliott Peranson. 1997. "The Effects of the Change in the NRMP Matching Algorithm." *Journal of the American Medical Association*, 278(9): 729–732.

Roth, Alvin and Elliot Peranson. 1999. "The Redesign of the Matching Market for American Physicians: Some Engineering Aspects of Economic Design." *American Economic Review,* 89(4): 748-780.

Stiglitz, Joseph E., Amartya Sen, and Jean-Paul Fitoussi. 2009. *Report by the Commission on the Measurement of Economic Performance and Social Progress*. www.stiglitz-sen-fitoussi.fr

Tversky, Amos, and Dale Griffin. 1991. "Endowments and Contrast in Judgments of Well-Being." In *Strategy and Choice*, ed. Richard J. Zeckhauser. Cambridge, MA: MIT Press. Reprinted in *Choices, Values, and Frames*, ed. Kahneman, Daniel, and Amos Tversky. Cambridge, UK: Cambridge University Press.

Van den Berg, Barnard and Ada Ferrer-i-Carbonell. 2007. "Monetary Valuation of Informal Care: The Well-Being Valuation Method." *Health Economics,* 16: 1227–1244.

Van Praag, Bernard and Barbara Baarsma. 2005. "Using Happiness Surveys to Value Intangibles: The Case of Airport Noise." *Economic Journal,* 115(500): 224–246.

Table 1: Main SWB and Residency Attribute Survey Questions

Variable label	Question prompt (beginning "On a scale from 1 to 100, …")
Happiness during residency	…how happy do you think you would feel on a typical day during this residency?
Life satisfaction during residency	…how satisfied do you think you would be with your life as a whole while attending this residency?
Ladder	…where 1 is "worst possible life for you" and 100 is "best possible life for you" where do you think the residency would put you?
Residency prestige and status	…how would you rate the prestige and status associated with this residency?
Social life during residency	...what would you expect the quality of your social life to be during this residency?
Desirability of location	…taking into account city quality and access to family and friends, how desirable do you find the location of this residency?
Anxiety during residency	…how anxious do you think you would feel on a typical day during this residency?
Worthwhile life during residency	…to what extent do you think your life would seem worthwhile during this residency?
Stress during residency	…how stressed do you think you would feel on a typical day during this residency?
Future career prospects	…how would you rate your future career prospects and future employment opportunities if you get matched with this residency?
Control over life	…how do you expect this residency to affect your control over your life?
Desirable for significant other	…how desirable is this residency for your spouse or significant other?

Table 2: Marginal Utility Estimates of Residency Attributes: Choice vs. Anticipated SWB

	(1) Choice	(2) Happiness during residency	(3) Life satisfaction during residency	(4) Ladder	(5) 4-interval-happiness index	(6) 3-SWB-measure index	(7) 6-SWB-question index
Residency prestige and status	2.5*** (0.3)	0.0 (0.3)	0.7* (0.3)	0.9** (0.4)	0.3 (0.4)	0.8** (0.3)	1.1** (0.4)
Social life during residency	1.6*** (0.3)	3.3*** (0.4)	2.7*** (0.4)	3.2*** (0.4)	2.6*** (0.4)	3.6*** (0.3)	3.5*** (0.5)
Desirability of location	1.7*** (0.2)	0.4* (0.2)	1.7*** (0.3)	1.9*** (0.3)	0.5* (0.3)	1.9*** (0.2)	1.6*** (0.3)
Anxiety during residency	-0.3 (0.3)	-1.3*** (0.3)	-0.5 (0.4)	-0.8** (0.3)	-1.8*** (0.4)	-0.9*** (0.3)	-1.4*** (0.4)
Worthwhile life during residency	4.4*** (0.5)	6.3*** (0.6)	7.0*** (0.6)	6.4*** (0.6)	5.9*** (0.7)	6.5*** (0.6)	6.9*** (0.8)
Stress during residency	-0.1 (0.3)	-1.0*** (0.4)	-0.7** (0.4)	-0.6* (0.3)	0.5 (0.4)	-0.7** (0.3)	0.0 (0.4)
Future career prospects	3.2*** (0.5)	0.9* (0.5)	1.8*** (0.5)	3.0*** (0.5)	1.2** (0.6)	2.6*** (0.5)	2.8*** (0.7)
Control over life	0.4 (0.3)	0.9** (0.3)	0.4 (0.3)	0.4 (0.3)	1.0** (0.4)	0.4 (0.3)	1.5*** (0.4)
Desirable for significant other	2.6*** (0.3)	0.5* (0.3)	0.7*** (0.3)	1.0*** (0.3)	0.3 (0.3)	1.2*** (0.2)	0.9*** (0.3)
# Observations	2169	2167	2169	2168	1591	2166	1590
McKelvey & Zavoina R^2, within variance only	0.40	0.28	0.35	0.40	0.24	0.41	0.42
# Students	557	557	557	557	540	557	540
Joint significance of differences with choice coefficients		0.000	0.000	0.000	0.000	0.000	0.000

Notes: Standard errors in parentheses. Rank-ordered logit regressions of either choice (column 1) or a SWB measure (columns 2–7) on residency attributes. Only ordinal information on the dependent variables is used. Columns 2–4 use the ordinal rankings implied by main three SWB measures. Columns 5–7 use the ordinal rankings implied by an optimal linear utility index, created by a first-stage rank-ordered logit regression of choice on the index components. All attribute ratings are divided by 100 before being included in the regression. Joint significance of the differences with choice coefficients (bottom row): *p*-value from a Wald test of the joint equality of all coefficients in the column with all coefficients in the choice column. * $p < .1$, ** $p < .05$, *** $p < .01$

Table 3: Normalized MRS Estimates of Residency Attributes: Choice vs. Anticipated SWB

	(1) Choice	(2) Happiness during residency	(3) Life satisfaction during residency	(4) Ladder	(5) 4-interval-happiness index	(6) 3-SWB-measure index	(7) 6-SWB-question index
Residency prestige and status	1.4***	0.0	0.4*	0.4**	0.2	0.4**	0.5**
	(0.2)	(0.2)	(0.2)	(0.2)	(0.3)	(0.2)	(0.2)
Social life during residency	0.8***	2.0***	1.5***	1.6***	1.7***	1.7***	1.6***
	(0.2)	(0.2)	(0.2)	(0.2)	(0.3)	(0.2)	(0.2)
Desirability of location	0.9***	0.3*	1.0***	0.9***	0.3*	0.9***	0.7***
	(0.1)	(0.2)	(0.1)	(0.1)	(0.2)	(0.1)	(0.1)
Anxiety during residency	-0.1	-0.8***	-0.3	-0.4**	-1.1***	-0.4***	-0.6***
	(0.2)	(0.2)	(0.2)	(0.2)	(0.2)	(0.2)	(0.2)
Worthwhile life during residency	2.4***	3.9***	3.9***	3.2***	3.7***	3.1***	3.2***
	(0.2)	(0.3)	(0.3)	(0.3)	(0.4)	(0.2)	(0.3)
Stress during residency	-0.1	-0.6***	-0.4**	-0.3*	0.3	-0.3**	0.0
	(0.2)	(0.2)	(0.2)	(0.2)	(0.3)	(0.2)	(0.2)
Future career prospects	1.7***	0.5*	1.0***	1.5***	0.8**	1.3***	1.3***
	(0.3)	(0.3)	(0.3)	(0.3)	(0.4)	(0.2)	(0.3)
Control over life	0.2	0.5***	0.2	0.2	0.6**	0.2	0.7***
	(0.2)	(0.2)	(0.2)	(0.2)	(0.3)	(0.1)	(0.2)
Desirable for significant other	1.4***	0.3*	0.4***	0.5***	0.2	0.6***	0.4***
	(0.1)	(0.2)	(0.1)	(0.1)	(0.2)	(0.1)	(0.1)
# Observations	2169	2167	2169	2168	1591	2166	1590
# Students	557	557	557	557	540	557	540
Joint significance of differences with choice coefficients		0.000	0.000	0.000	0.000	0.000	0.000

Notes: Delta-method standard errors in parentheses. Entries are coefficients from table 2, normalized by taking the ratio to the average absolute value of the nine coefficients in a column from table 2. Joint significance of the differences with choice entries (bottom row): *p*-value from a Wald test of the joint equality of all entries in the column with all entries in the choice column.* $p < .1$, ** $p < .05$, *** $p < .01$

Table 4: Weight Estimates for Multi-Measure Indices

	(1) Choice	(2) Choice	(3) Choice
Happiness during residency	4.5***	0.6	0.9
	(0.5)	(0.4)	(0.6)
Happiness in first 10 years	4.6***		3.5***
	(0.8)		(0.9)
Happiness in rest of career	2.1**		2.4***
	(0.9)		(0.9)
Happiness after retirement	1.2		2.0**
	(0.8)		(0.9)
Life satisfaction during residency		4.4***	3.9***
		(0.5)	(0.7)
Ladder		5.5***	5.4***
		(0.4)	(0.6)
# Observations	1609	2192	1607
McKelvey & Zavoina R^2, within variance only	0.13	0.31	0.31
# Students	544	561	544

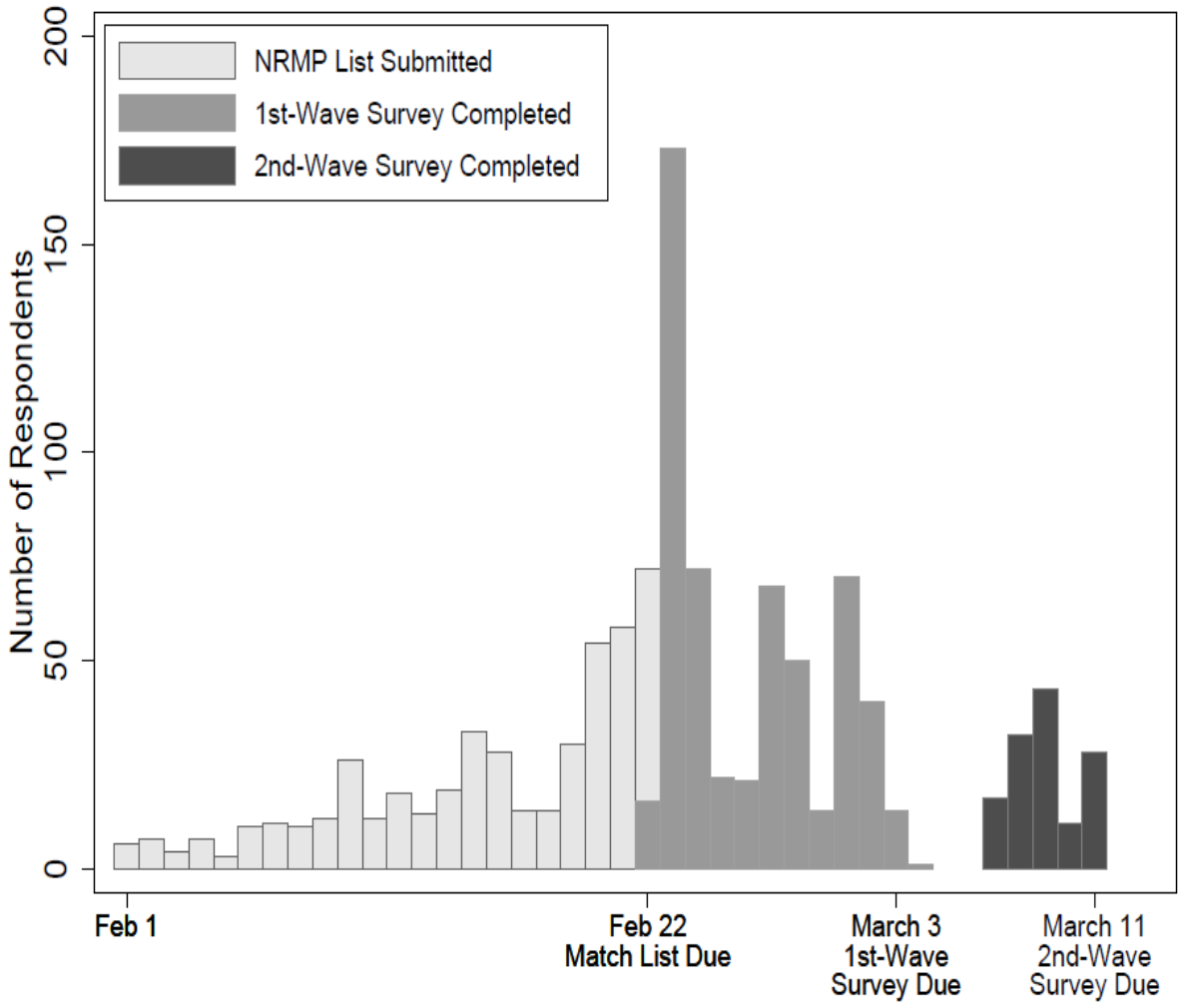
Notes: Standard errors in parentheses. Rank-ordered logit regressions of choice on SWB measures. All aspect ratings are divided by 100 prior to inclusion in the regression. Since future happiness measures are only elicited for three of the four ranked residencies, less data is available for conducting these regressions relative to those with only the primary SWB questions. However, restricting all three regressions to the same sample of 1607 observations has only minor impact on the coefficient estimates and R^2's (column 1's $R^2 = 0.14$; column 2's $R^2 = 0.27$). * $p < .1$, ** $p < .05$, *** $p < .01$

Table 5: Predicting Binary Choice from Anticipated-SWB and Attribute Questions

	(1) Preferred program has higher rating	(2) The two programs have same rating	(3) Preferred program has lower rating	(4) Correct prediction rate (1)/(100%-(2))	(5) # Pairwise program comparisons
Happiness during residency	52%	27%	21%	71%	3241
Life satisfaction during residency	59%	23%	18%	77%	3245
Ladder	65%	18%	17%	80%	3246
Residency prestige and status	56%	16%	28%	67%	3244
Social life during residency	52%	20%	28%	65%	3248
Desirability of location	61%	14%	25%	71%	3242
Anxiety during residency	38%	29%	33%	53%	3237
Worthwhile life during residency	44%	40%	16%	73%	3236
Stress during residency	40%	26%	34%	54%	3237
Future career prospects	49%	30%	21%	70%	3247
Control over life	40%	30%	30%	57%	3235
Desirable for significant other	65%	16%	19%	77%	2087
Average happiness in first 10 years	34%	53%	13%	72%	1603
Average happiness in rest of career	28%	56%	16%	64%	1603
Average happiness after retirement	22%	64%	14%	62%	1605
4-interval-happiness index	62%	10%	28%	69%	1592
3-SWB-measure index	75%	3%	22%	77%	3233
6-SWB-question index	76%	2%	22%	78%	1588
12-question index (3 SWB + 9 attribute)	81%	0%	19%	81%	3179
15-question index (6 SWB + 9 attribute)	81%	0%	19%	82%	1566

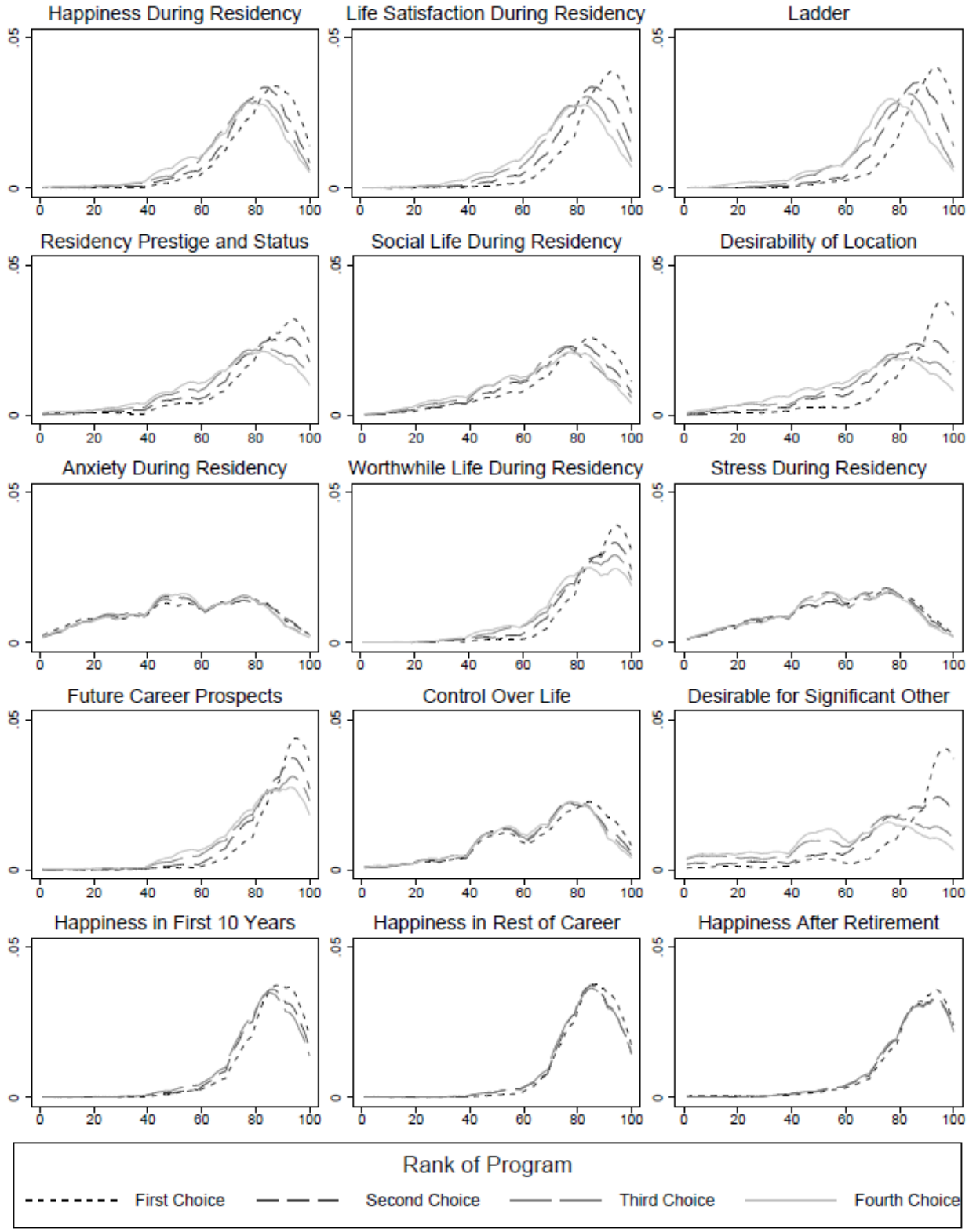
Notes: Based on only the ordinal ranking of the variable in each row. All six binary comparisons among the top four programs are considered. Columns 1–3 sum to 100% in each row. Column 4 reports the correct prediction rate in cases where a prediction is made; that is, excluding cases of indifference (column 2). Column 5 reports sample size.

Figure 1: Survey Response Timeline



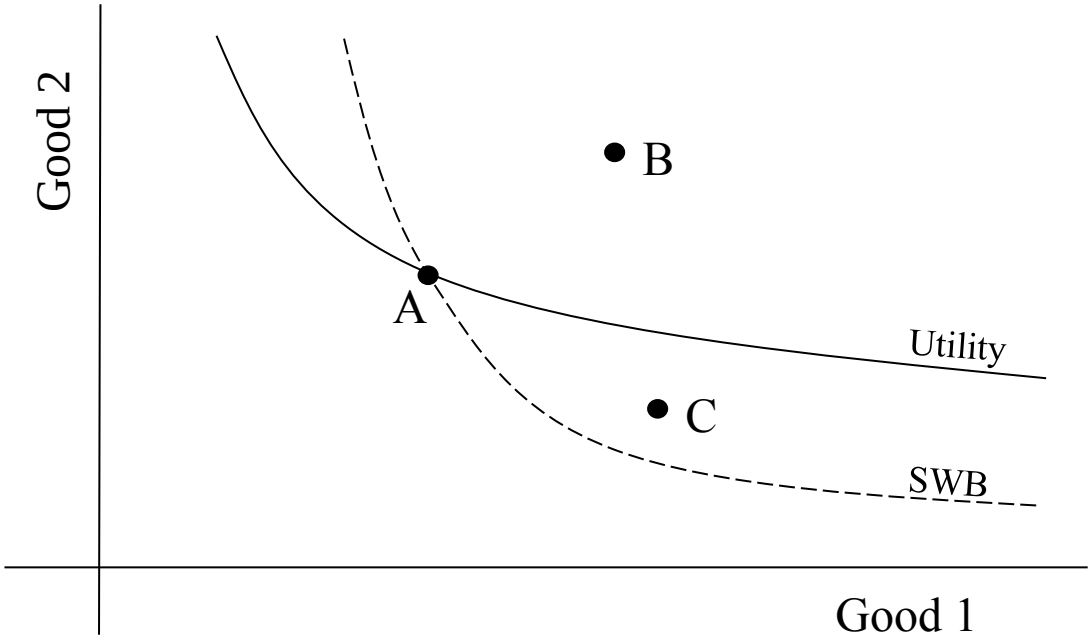
Notes: Frequency distribution of survey-response dates. NRMP submission and 1st-wave data are for the 561 respondents in our main sample. 2nd-wave data are for the 131 respondents in the main sample who completed the repeat survey. Each bar corresponds to one day. The 1st wave responses submitted on Feb 22nd occurred after 9pm, the deadline for match list submission.

Figure 2: Distributions of Variables by Program Rank



Notes: Based on 561 respondents in the main sample. Kernel density plots of residency attributes by preference order. (Epanechnikov; Bandwidth 5.)

Figure 3: Implications of Iso-Utility and Iso-SWB Curves for Ordinal Prediction



Notes: This figure illustrates the implications of different tradeoffs in revealed-preference utility and SWB for binary comparisons. The solid line represents an individual's iso-utility curve, while the dashed line represents her iso-SWB curve. When comparing option A to option B, the iso-utility curves and iso-SWB curves imply the same binary ordering. When comparing option A to option C, they differ: option C has higher SWB, but is less preferred. The figure suggests that with different implied tradeoffs, whether SWB data still yield ranking of choice options that coincides with preference ranking depends on the exact shapes of the iso-utility and iso-SWB curves, as well as on the orientation of the considered options relative to each other.