Stanford cepa | Center for Education Policy Analysis

CEPA Working Paper No. 16-06

Welfare Adjusted Scale Score: Method Toward the Development of an Equal-Interval Welfare Scale

AUTHORS

Ken Shores
Stanford University

ABSTRACT

It is common to question the equal-interval assumptions of most academic scale scores. Even if interval assumptions hold, it is problematic to assume equal-interval distances with respect to benefits. For example, equivalent gains at the top and bottom of the distribution are unlikely to yield equivalent welfare returns. I develop a method to estimate the welfare returns to academic achievement directly, by making use of established methodologies in health economics. Using performance level descriptors and achievement data from the National Assessment of Educational Progress Long Term Trends, I estimate a random utility model to construct a welfare-adjusted equal interval scale. I then show that welfare returns to achievement are non-linear, convex and lead to different inferences regarding achievement gap trends.

VERSION

February 2016

Suggested citation: Shores, K. (2016). Welfare Adjusted Scale Score: Method Toward the Development of an Equal-Interval Welfare Scale (CEPA Working Paper No.16-06). Retrieved from Stanford Center for Education Policy Analysis: http://cepa.stanford.edu/wp16-06

Welfare Adjusted Scale Score: Method Toward the Development of an Equal-Interval Welfare Scale *

Kenneth A. Shores

Stanford University

February 11, 2016

Abstract

It is common to question the equal-interval assumptions of most academic scale scores. Even if interval assumptions hold, it is problematic to assume equal-interval distances with respect to benefits. For example, equivalent gains at the top and bottom of the distribution are unlikely to yield equivalent welfare returns. I develop a method to estimate the welfare returns to academic achievement directly, by making use of established methodologies in health economics. Using performance level descriptors and achievement data from the National Assessment of of Educational Progress Long-Term Trends, I estimate a random utility model to construct a welfare-adjusted equal-interval scale. I then show that welfare returns to achievement are non-linear, convex and lead to different inferences regarding achievement gap trends.

Keywords: Achievement, Achievement Gaps, Cost-Effectiveness Analysis, Equal-Interval, Random Utility Model, Scaling, Welfare

DRAFT. Please do not cite without permission.

*Direct correspondence to Shores (kshores@stanford.edu); 520 Galvez Mall, CEPA 5th Floor, Stanford, CA 94305. I wish to acknowledge generous support from the Institute for Educational Sciences (Grant #R305B090016) and National Academy of Education/Spencer Dissertation Fellowship for funding this work. The author especially thanks Ed Haertel, Andrew Ho, Susanna Loeb, Sean Reardon, Paul Sniderman and Mike Tomz for comments and suggestions. The author also thanks Ben Shear for helpful conversation as well as the participants in Stanford's Laboratory for the Study of American Values. All errors are my own.

I. Introduction

The use of academic scale scores for program evaluation or population description requires that equal-interval assumptions are met. In fact, any aggregation of test scores requires that distances between units are equivalent throughout the range of the scale score. Item Response Theory (IRT) offers a set of models that can, under certain conditions, estimate a scale with equal-interval properties (see Ballou, 2009 for review). Nevertheless, the assumptions that allow us to assume equal-interval properties are not often tested. Current research suggests that equal-interval assumptions are problematic. Domingue and, in a working paper, Nielsen have developed methods for testing whether the equal-interval assumptions are plausibly met for some common academic assessments and find that these assumptions are not (Domingue, 2014; Neilsen, 2015).

Given the skepticism about the interval properties of most test scores, other researchers simply assume that any given scale is but one among many monotone transformations of a latent scale. Given this agnosticism, Cunha and Heckman (2008) and Cunha, Heckman and Schennach (2010) propose a scale transformation that anchors the original scale to adult earnings, a distribution that is assumed to have equal-interval properties. The transformed scale score is then used to estimate production functions for cognitive development. Relying on similar assumptions about the flexibility of scale transformations, Bond and Lang (2013a; 2013b) subject a scale score to a variety of monotone transformations according to an algorithmic objective function that maximizes and minimizes changes in the white-black achievement gap. The authors find that inferences about gap changes are, not surprisingly, sensitive to these scale transformations.

Indexing achievement scores to future earnings is not without its own problems. First, income can be scaled to reflect either real dollars or the natural logarithm. Whether the outcome is measured in real dollars or log-transformed dollars has been shown to affect inferences (Lee and Solon, 2011; Solon, Haider and Woolridge, 2015). Furthermore, income, however scaled, is not an inclusive indicator of benefit. Plenty of non-pecuniary benefits can be attributed to achievement, such as the simple pleasures one gets from being numerate and literate.

The approach that I take here is to assume an equal-interval scale in the metric of achievement and estimate a new scale that will be equal-interval in the metric of welfare. The method I propose estimates utility for a set of 10 "achievement states", where an achievement state corresponds to a performance level descriptor for reading and math taken from the National Assessment of Educational Progress Long-Term Trend (NAEP LTT), and utility corresponds to how much better, all things considered, a sample of Americans believe a person's life will be for a given achievement state.[1] Because the NAEP uses a scale anchoring process to link scale scores to performance level descriptors, it is possible to build a data set with three variables and 10 observations: a vector of performance level descriptors, the corresponding scale scores, and the estimated utilities.

In order to link the discrete utility estimates to the full range of continuous scale scores, I use piece-wise monotone cubic interpolation (MCI) to link individual NAEP scores to a utility value.[2] We now have a scale score that is equal-interval with respect to welfare, as long as the equal-interval assumptions (with respect to ability) of the original scale score hold and that the performance level descriptors are appropriately mapped to scale score values. As a demonstration of the usefulness of a welfare scale, I take a repeated cross-sectional panel of student test scores from the NAEP-LTT and re-scale them according to the method outlined above. I show that inferences about changes in achievement and achievement gaps over time and age are sensitive to the choice of scale.

This method is similar in concept to methods commonly used in health care research. In health economics, effect sizes are, in many cases, given in the metric of a Quality Adjusted Life Year (QALY) (see Drummond, 2005 and Whitehead and Shehzad, 2010 for review), where the QALY metric is used to make comparisons between different 'health states' (where health states are the analogue to achievement states taken from performance level descriptors) in health care production functions for purposes of cost-effectiveness analysis. As an example, consider a medical intervention that improves mobility by 2-units and another that

[1]Performance level descriptors for reading can be found online here: https://nces.ed.gov/nationsreportcard/ltt/reading-descriptions.aspx; math here: https://nces.ed.gov/nationsreportcard/ltt/math-descriptions.aspx. 5 performance level descriptors are available for both reading and math, for a total of 10 descriptors. See Data section for details.

[2]MCI is implemented according to Fritsch and Carlson, 1980.

reduces pain by 3-units. Holding costs constant, researchers, insurance companies and policy makers are interested in determining which of the two interventions should be pursued. The QALY-metric puts discrete health outcomes on a common utility scale, making comparisons possible. In addition to being used for making between health state comparisons (e.g., mobility against pain), QALY-scales can be used for making within health state comparisons (e.g., completely immobile against able to walk without assistance).[3]

A scale that could be used in educational settings for making analogous between- and within-state comparisons would be useful insofar as we wish to understand whether and in what cases changes in achievement are more or less important. For instance, current program evaluations leave fundamental questions unanswerable. Holding costs constant, if one intervention raises math scores 10-units and another raises reading scores 10-units, we lack an outcome variable that adjudicates between the two interventions. Likewise, if one intervention raises math scores 10-units at the low end of the scale and another intervention raises math scores 10-units at the high end of the scale, current practice fails to distinguish between these two results. The method I describe and implement is one solution for resolving this uncertainty.

In order to head-off criticism, I emphasize that the scaling procedure described here is intended as proof of concept. It provides one way for constructing an equal-interval scale that can be used for descriptive and evaluative purposes. Moreover, the procedure brings some fundamental questions of measurement into sharp relief. Consider:

1 To what outcome should scale scores be indexed? In this paper, I present respondents with questions, asking them to determine which description of math and reading is more important for an "all things considered" better life. Other indices are available, such as income, civics engagement, or health outcomes. Cunha and colleagues (2008; 2010) index a child's test score to the child's future earnings, using a factor loading technique to weight the achievement distribution as a function of how well it predicts earnings. Such a technique is not a panacea however. First, the factor loading method is an *ad hoc* scaling technique. Second, earnings connote their own scaling assumptions, e.g.,

[3]The Eq-5d, for example, is one of the more commonly used metrics and provides three descriptions of mobility states, three descriptions of pain states, as well as three other health states. Utility scores are estimated for each health state, allowing for between and within health state comparisons (Oppe, Devlin and Szende, 2007).

should the scale be log transformed? Finally, linking achievement to earnings ignores the academic capabilities captured by the scale score.

2 Whose preferences for achievement should be included in the index? The sample of respondents included in this essay are mostly college educated. In the survey experiment, respondents are asked which state of achievement is more important for a good life. If respondents have no understanding of what a high level of numeracy or literacy feels like or entails, they will struggle to respond to the question. This suggests college educated respondents are appropriate. Nevertheless, it is likely important that an index of benefit captures the preferences of everyone. How to include all respondents in an exercise for which some may lack the cognitive capacity to participate is a difficult question. Note that the question does not apply to achievement alone: whether the poor can predict how much they would prefer being non-poor (and vice-versa) seems similarly opaque. Whether a healthy person can predict how much they are pain averse has a similar problem.[4]

3 How should the index balance individual and social benefits? The approach used here measures preferences for individual benefits to achievement states. Such an approach ignores other distributive concerns, such as equity. It is known that survey respondents may be indifferent between a 1-unit change at the bottom and top of a scale when comparing between two persons, but when respondents are asked which of the two persons should receive treatment in a group of persons, they will choose to give treatment to the person whose health is at the bottom of the scale (Nord, 1999). This suggests individuals value relative differences (Otsuka and Voorhoeve, 2009). How such equity concerns, and other social values, should be included in the index is an important question.

4 How should time be modeled in the elicitation and estimation of the utility value? The method used here (and the one that is commonly employed in health economics) is to present respondents with a cross-sectional preference: "Person A has characteristics X and Person B has characteristics X'. Who is better?" In health, these characteristics are fixed by specifying that the health state will persist for t-years, whereas achievement states are naturally assumed to change over time as students learn. Moreover, individuals may have different preferences for achievement growth than they do for achievement states. Linking a preference for achievement change to a student's scale score is complicated because our current measures of achievement only provide cross-sectional information about the student's abilities.[5]

[4]Dolan and Kahneman call this distinction experience versus decision utility.

[5]See Lipscomb, et al., 2009 for a review of this and other issues related to time in the health landscape.

These questions are a current source of debate in health economics and philosophy,and are likely to continue to be debated.[6] Questions like these are currently neglected in most education policy evaluation, or the answers supplied to the questions are left as unstated assumptions.

The paper proceeds as follows: I begin by providing an overview of the survey design and the data used for analysis. I then describe the theoretical model that motivates the analysis and the econometric model that will be used for empirical estimation. The first set of results I show describe utility values for the different achievement states. Interpolation techniques are described that connect NAEP scores to utility values for the full distribution of NAEP data. As an example, I estimate white-black achievement gaps using the original and welfare-adjusted NAEP data and show that inferences about gap trends are sensitive to scale selection.

II. Survey design

The survey design has two components. The first is a ranking exercise, in which three out of five reading or three out of five math descriptors are randomly selected and respondents are asked to rank these descriptors in order of difficulty. Reading and math ability descriptions are taken from the NAEP-LTT performance level descriptors, described below. The purpose of this ranking exercise is two-fold: to prime respondents so that they recognize these descriptors are ordinally ranked, and to screen respondents who cannot (or will not) rank descriptors correctly. Figures II through III display the ranking and choice tasks as they appeared in the experiment.

The ranking exercise is followed by a choice-based conjoint design (often times referred to as a discrete choice experiment) to obtain utility values for different math and reading descriptors. Choice-based conjoint designs are widespread in health and public economics, marketing research, and have become increasingly common in political science (for examples

[6]See Daniels, 1985 for a philosophical view of justice in the provision of health care, as well as Nord, 1999 who offers a mixture of economics and philosophy in his evaluation of the QALY metric.

in health and public economics, see De Bekker-Grob, et al., 2012 and McFadden, 2001, respectively; in marketing, see McFadden, 1986; in political science see Hainmueller and Hopkins, 2014). In the experiment, respondents are provided with a description of two individuals (Person A and Person B) who are alike in all respects, except that they differ in their math and reading abilities. Respondents are asked to determine which bundle of math and reading abilities between Persons A and B will lead to an "all things considered" better life. After being presented with the reading and math profiles, the respondent is forced to make a choice between Persons A and B. The response is coded dichotomously, 1 if Person A or B was chosen and 0 otherwise. Each respondent is given only one choice task.[7]

The purpose of the choice task is for respondent to make *interval* comparisons between Persons A and B with respect to welfare. As an example, consider a choice task where Person A has reading ability equal to 5 and math ability equal to 2, while Person B has reading and math abilities equal to 3.[8] Effectively, the respondent is being asked to make a trade between 2 units of reading for 1 unit of math. Whether respondents, on average, choose Person A over B will depend on how much they value reading relative to math, and, importantly, how much they value math gains at the bottom of the distribution relative to reading losses at the top. To see this, consider an alternative choice task where Person A has reading ability 4 and math ability 1 and Person B has reading and math abilities 2. Here, the reading and math abilities of Persons A and B have been shifted down equally, but respondents may not make the same selections, since a change in reading from 5 to 4 need not be equivalent to a change in reading from 4 to 3. This exercise formally tests whether respondents' preferences are, indeed, equal interval with respect to welfare. Depending on how respondents on average weight these different trades will determine the relative concavity of the welfare-adjusted scale score.

[Insert Figure I Here]

[7]More than one choice task is of course possible, requiring that standard errors be clustered at the respondent level. The decision to offer respondents only one choice task was motivated by a reduction in cognitive load, as performance descriptors are text heavy, as well as the fact that the marginal survey cost using Amazon's MTurk suite are relatively low.

[8]Where ability level 5 corresponds to highest performance level descriptor on the NAEP, and so on. Respondents are not asked to make trades regarding integer values of the NAEP but are instead presented with textual descriptions of reading and math abilities commensurate with integer scores. See Scale Anchoring section below.

[Insert Figure II Here]

[Insert Figure III Here]

Finally, note that Figure II explicitly states the age of Persons A and B. Because the performance level descriptors from the NAEP-LTT pertain to students at the ages of 9, 13 and 17, and because individual scale scores are available for students at those ages, I randomly assign one of three ages (9, 13, 17) to each choice task. The purpose of this additional randomization is to test the sensitivity of preferences for achievement bundles to age. For example, respondents may value gains in reading and math at the low end of the distribution for persons aged 9 more than they value equivalent gains for persons aged 17. Randomly assigning age will allow me to test this hypothesis.

II.A. Math and reading descriptors and scale scores

The choice task described above uses performance level descriptors to connote reading and math ability levels. In order to construct a data set with performance level descriptors, utility values, and scale scores, it is necessary that these performance level descriptors (and their estimated utilities) can be plausibly linked to scale scores. The plausibility of this linking is defended below, but it is natural to wonder why performance level descriptors are needed at all. Why not ask respondents to make trades using the scale scores themselves?

There are two problems with such an approach. To the first, I hope it is evident that in order to estimate an interval scale with respect to welfare it is important not to conflate the welfare scale with the original scale that describes ability. The purpose of the choice task is to allow respondents to decide for themselves the interval distances with respect to value between, say, reading units 1 and 2 and units 3 and 4, and so on. A Rasch or IRT model might estimate equal-interval distances between these units, but respondents are being asked to decide whether these distances are equal in another dimension, that of welfare. The second problem is that a scale score decoupled from a performance level descriptor connotes no meaningful information to the respondent. Any scale can be linearly transformed, and determining how much 5 units is worth relative to 4 units, or 500 relative to 400 is not possible. For these reasons, it is necessary to provide respondents with a richer descriptor of

what performance looks like, and then link the performance-level descriptor back to a scale score.

II.B. Linking NAEP descriptors to scale scores

I now turn to the question of whether or not performance-level descriptors can be plausibly mapped onto scale scores. One of the goals of academic measurement, dating back to at least 1963, is to provide criterion-referenced interpretations of scale scores—in other words, to be able to provide descriptions of what students know and can do in an academic domain (Mullis and Jenkins, 1988). The process by which the NAEP links performance level descriptors to scale scores is called scale anchoring. Scale anchoring attempts to provide a context for understanding the level of performance defined by the specific test items that are likely to be answered correctly by students (Lissitz and Bourqe, 1995).

Anchor levels are determined by a combination of statistical and judgmental processes. For the NAEP, an IRT model is used to estimate an ability score, θ, for each student, bounded between 0 and 500. The equidistant points 150, 200, 250, 300 and 350 are then selected from the scale.[9] Test items from the assessment are then selected and categorized according to whether or not the item discriminates between students with different scale scores. For example, an item will be categorized as a "150 level item" if (a) 65 percent of students scoring at or around 150 answered the item correctly; (b) 30 percent of students or fewer scoring below 150 answered it correctly; (c) no more than 50 percent of students scoring below 150 answered it correctly; and (d) a sufficient number of students responded to the item. With this procedure, a large number of items can be categorized as being "150 level items", "200 level items", and so on. This completes the statistical part of the process. The judgmental part of the process occurs when teams of curriculum and content specialists from the respective domains (i.e., reading and math) are asked to describe the kinds of academic competencies reflected in the categorized items. Specialists meet in teams and form a consensus about what these items signal.

[9]Very few students score in the tails of the scale score distribution, and for this reason the selected points of interest ignore those regions.

The final result is a set of performance level descriptors that characterize what students know and can do as defined by test performance on selected items. Scale scores are empirically determined, anchor items are empirically identified, and anchor descriptions are provided by expert judgment (see Beaton and Allen, 1992; Beaton and Johnson, 1992; Lissitz and Bourqe, 1995 for full description of the scale anchoring process).[10]

There are problems with this procedure. Lissitz and Bourque describe the anchor item selection process as "low inference" and the descriptive process as "high inference." The key issue revolves around whether the descriptors are overly uni-dimensional. Not all items can be empirically anchored to different ability levels, leaving open the possibility that the anchored items are too narrow. While experts construct uni-dimensional descriptions of anchor items, other descriptions cannot be ruled out. Moreover, the performance level descriptors collapse across different sub-scales, glossing over multi-dimemsionality that is present even in the empirical data. Finally, even though equidistant anchor levels are selected, if the equal-interval assumptions of the scale score are not met, then the descriptors will likewise not be equal-interval scaled.

Despite these concerns, anchoring in this way is the most widely used technique for providing descriptions of what students know and can do at different points across the scale. Given how widely these benchmarks are used in classrooms and policy discussions, it is at least plausible to suggest that the performance descriptors used in this survey experiment can be mapped to specific scale scores. The performance level descriptors for reading and math are described in Tables 1 and 2 below. The entire performance level descriptor is used in the choice-based conjoint experiment.

[Insert Table I Here]

[Insert Table II Here]

[10]Performance level descriptors differ from standards setting or achievement level descriptors. Standards setting practices begin with a set of skills that experts believe correspond to proficiency levels. For instance, it might be asserted that a 4th grade student is proficient in reading if that 4th grade student can read chapter books for comprehension. Experts then work through the test items and determine subjectively what percent of students would answer the item correctly, if the student was proficient in reading. This stands in stark contrast to the anchoring procedure described here, as the items are not categorized according to a statistical procedure and given subjective analysis ex post, but instead are categorized exclusively according to a judgmental procedure.

II.C. Data collection

Utility values are estimated from survey respondents. Respondents are drawn from the United States during the period of June and July, 2015. Participants were enrolled using Amazon's Mechanical Turk software suite and the survey was administered using Qualtrics. Respondents were offered $0.35 to participate in the survey, equivalent to about $6.00 per hour, and the study was administered with IRB approval. In total, 2351 respondents participated. According to self-reports, respondents were primarily college educated (78 percent), white (73 percent) and balanced by gender (48 percent male, 52 percent female). Mechanical Turk populations, while not representative of the national population on observables, have been shown to have nationally representative preferences with respect to certain stimuli, such as responsiveness to information about income distributions (Kuziemko, Norton, and Saez, 2015) and risk aversion (Buhrmester, Kwang and Gosling, 2011).[11]

III. Econometric framework

In this section I describe the modeling approach I use to estimate utility values for each of the math and reading performance level descriptors. The model uses the logistic likelihood function to provide point estimates for reading and math performance level descriptors at levels 150, 200, 250, 300 and 350. Point estimates for reading and math performance level descriptors can be interpreted as the log likelihood that respondent i chose Person A (profile 1) with reading and math characteristics θ_{sl}, where s indexes subject (reading or math) and l indexes performance level (150, 200, 250, 300, 350).

Formally, the data are structured so that there is one row of observation for each survey respondent i. A response variable is coded 1 if respondent chose Person A (profile 1); 0 otherwise–that is, if the perceived utility of Person A exceeded the perceived utility of Person B. The pairwise offerings presented to each respondent are coded as indicator variables. For example, if respondent i compared Person A, who had Reading 150 Math 300 (Reading

[11] Pilot studies took place over the months of September, 2014 to June, 2015. Development of the survey design took place in Stanford's Laboratory for the Study of American Values.

1, Math 4) and Person B, who had reading 200 Math 250 (Reading 2, Math 3), the indicator variables Read1a, Math4a, Read2b and Math3b would be coded 1; all other indicator variables (Read2a through Read5a, Math1a-Math3a and Math5a, etc.) are coded 0. These ones and zeroes mark the choice set available to the respondent.

Thus, the probability that respondent i chose Person A is:

$$(1) \quad Pr(ChooseA) = f(U_{ia} + \varepsilon_{ia} > U_{ib} + \varepsilon_{ib}),$$

$$(2) \quad = f(U_{ia} - U_{ib} + \varepsilon_{ia} - \varepsilon_{ib} > 0),$$

This expression says that the probability of choosing Person A is a function of an individual's observed utility for Persons A and B plus a random component ε_{ij}. Respondents choose A when they perceive more utility for A than B, or when the difference in utility between Persons A and B is greater than zero.

If we assume that the errors have a logistic distribution, then we can specify the model such that:

$$(3) \quad Pr(ChooseA) = 1 + \frac{1}{e^{-(U_{ia}-U_{ib})}} + \varepsilon_{ij}; \varepsilon_{ij} = \varepsilon_{ia} - \varepsilon_{ib}$$

$$(4) \quad = 1 + \frac{1}{e^{U_{ib}-U_{ia}}} + \varepsilon_{ij}$$

We simplify by taking logs and get:

$$(5) \quad Ln\frac{Pr(ChooseA)}{Pr(ChooseB)} = U_{ib} - U_{ia} + \mu_{ij}$$

So far we have only shown that the log odds of choosing Person A over B will be a function of how much the utility attributed to Person A exceeds the utility attributed to Person B. We also know that Persons A and B have characteristics. Substituting, we get:

$$(6) \quad U_{ib} = Math_{ib} + Read_{ib}; U_{ia} = Math_{ia} + Read_{ia}$$

$$(7) \quad Ln\frac{Pr(ChooseA)}{Pr(ChooseB)} = (Math_{ib} - Math_{ia}) + (Read_{ib} - Read_{ia}) + \mu_{ij}$$

This expression says that the log odds of choosing Person A over B will be a function of how much Person A's math and reading abilities ($Math_{ia}$ and $Read_{ia}$, respectively) are preferred over Person B's math and reading abilities ($Math_{ib}$ and $Read_{ib}$, respectively).

Let $\theta_{sl} = Math_{ib} - Math_{ia}$ or $Read_{ib} - Read_{ia}$ for the full vector of Math and Reading pairwise offerings made available to all respondents. Then, the model can be estimated with the equation:

$$(8) \quad Ln\frac{Pr(ChooseA)}{Pr(ChooseB)} = \alpha + \sum_{s=1}^{2}\sum_{l=2}^{5}\theta_{sli} + \mu_{sli}$$

Where s indexes subjects (reading and math), l indexes levels (200, 250, 300, 350 and 150 for both subjects is jointly estimated by the constant α), and i indexes respondents. This model estimates a total of 8 parameters plus a constant. Standard errors are clustered to account for heteroskedasticity.

Previously, I noted that the ages 9, 13 and 17 were randomly assigned to respondents, in order to test whether respondent preferences for different parts of the reading and math distributions varied by the supplied ages of Persons A and B. These age terms can be introduced in the model as interactions:

$$(9) \quad Ln\frac{Pr(ChooseA)}{Pr(ChooseB)} = \alpha + \delta_a \times (\sum_{s=1}^{2}\sum_{l=2}^{5}\theta_{sli}) + \mu_{sli}$$

where δ_a is an age fixed effect, thus giving 24 total parameters estimated (8 reading and math x 3 age terms) and a constant.

IV. Results

I now turn to results. The survey consisted of both a ranking and a choice exercise. The ranking exercise was included to determine whether respondents could and did understand that the performance level descriptors provided increasingly sophisticated descriptions of reading and math abilities. I begin by showing percents of respondents ranking performance

level descriptors correctly in terms of difficulty. A majority of respondents are able to rank these descriptors correctly, suggesting that they understand the descriptors connote ordinal information in terms of ability.

I then turn to point estimates from logistic linear regression models. I show point estimates for two sets of models: age-interaction models (for ages 9, 13 and 17) are shown along with models that estimate the weighted average across age. These allow us to see whether age-interactions meaningfully change respondent behavior. Three interpolation schemes are considered and monotonic cubic interpolation (MCI) according to Fritsch and Carlson (1980) is selected.

With an interpolation scheme in place, I have a full range of data for both the original scale score and the estimated welfare scale. As a descriptive application, I show trends in the white-black achievement gap, defined as the difference in mean white and black scores, for the original and adjusted scales. NAEP scores are fairly stable across time but change substantially as students age.[12] Test scores are available for a random sampling of students at ages 9 and 17 every 8 years for six cohorts in math and reading, allowing for description of achievement growth as students age across various cohorts. I conclude by showing gap trends across age for various cohorts using both scales.

IV.A. Ordinal ranking exercise

Respondents first participated in a ranking exercise in which they were randomly assigned 3 of 5 reading or 3 of 5 math performance level descriptors (an example of the exercise is shown in Figure I). Only three descriptors were randomly drawn in order to simplify the ranking task. There are 10 possible reading and math bundles for which there are no ties randomly assigned to respondents, when a tie is defined as respondent being randomly assigned one or more equivalent reading or math performance level descriptors.[13] Among non-ties, the probability of being assigned any one reading or math performance level descriptor is uniformly

[12]The NAEP-LTT is vertically scaled, meaning students at different ages are exposed to an overlapping subset of test items. See Beaton and Swick (1992) and Haertel (1991) for discussion.

[13]Ties are excluded because the exercise is made radically simpler when ranking only two unique sets of descriptors.

distributed. There are 10 possible non-tying combinations of performance level descriptors: 123, 124, 125, 134, 135, 145, 234, 235, 245, 345 (where 1=150, 2=200, 3=250, etc.). Likewise, the distribution of descriptor combinations is uniformly distributed.

Uniformity allows for independent point estimates of each subject-level descriptor. However, independent estimates of the effect of being assigned a performance level descriptor on the probability of ranking that descriptor correctly are available only if descriptor combinations are equally difficult. For example, some descriptor combinations will, by chance, assign respondents combinations of descriptors that are further spaced than other combinations (e.g., 135 is futher spaced than 234 or 125). If correctly ranking is easier when descriptors are further apart (e.g., 135 is easier than 234 or 125), and if some performance level descriptors are more commonly found in these more easily ranked combinations, then independent estimates of each performance level will be biased. To test for this, I construct three indicator variables (Distance 100, Distance 150, and Distance 200) indicating the cumulative distance between the three performance descriptors. For example, the indicator Distance 100 will be coded 1 if the three descriptors were 150, 200, 250 (distance is 50 between 150 and 200 and 50 between 200 and 250 for a total of 100); 0 otherwise. Distance 150 and 200 are coded similarly.

The data are structured such that there are three observations per respondent. Each row corresponds to the subject s and level l randomly shown to the respondent i. If the respondent ranked the item correctly, it is coded 1; 0 otherwise. In total, a respondent may rank 0,1 or 3 descriptors correctly (mis-ranking one descriptor necessarily results two or more descriptors mis-ranked). I estimate two regression models:

$$(10) \quad Rank_{sli} = \theta_{sli} + \mu_{sli}$$

$$(11) \quad = \delta_d \times (\theta_{sli}) + \mu_{sli}$$

Here, s indexes subject, l indexes level and i indexes respondent. Each model is run separately for math and reading, for a total of four estimations. The dependent variable $Rank_{sli}$ is coded as 0 or 1 depending on whether the respondent ranked correctly; indicator variables θ_{sli} indicate the linear probability that respondents ranked performance levels 1 through 5

correctly, and the interaction term δ_d indicates the proportion of respondents ranking θ_{sli} correctly when they were offered three-descriptor combinations with distances equal to 100, 150 or 200. Point estimates for δ_d interactions are relative to $d = 100$. Standard errors are clustered at the respondent level to account for intra-respondent correlation. Results are reported in Table III.

[Insert Table III Here]

The main effects coefficients (Levels 150 through 350, indicated by column header "Mean") indicate the proportion of respondents ranking that performance level descriptor correctly. Here we see that, for the most part, respondents were successful at ranking the descriptors. Percents correct range between 63 percent to 79 percent depending on subject and level. There is not an obvious pattern between subjects and levels with respect to how effectively respondents ranked.

The interaction terms confirm the hypothesis that additional space between performance level descriptors improves ranking competence. Relative to when cumulative distance is 100 (the smallest possible distance among non-ties), distances at 150 and 200 are nearly always higher (math, level 300, distance 200) and generally significant.[14] Overall, respondents ranked reading and math descriptors correctly 63 to 79 percent and 64 to 72 percent of the time, respectively. Whether respondents rank incorrectly on account of negligence or genuine confusion is unknown.

Correct rank ordering of performance level descriptors is relevant to the utility model because the model assumes monotonicity of consumer choice preferences. The monotonicity assumption is simply that respondents should choose higher levels of reading or math, all else constant. That is, for example, if respondent i faces a choice task k in which Person A has Reading and Math 250 and Person B has Reading 250 and Math 300, respondent i must choose Person B. In health economics, where choice-based conjoint designs are common and assumptions of monotonicity are likewise required, there is no consensus on best practices for when respondents "choose badly." I follow current practices and delete observations for

[14]The interaction terms do not average to the main mean effect because the Distance 150 terms are approximately 1.3 times more prevalent than either the 100 or 200 terms.

which respondent choices violate monotonicity assumptions.[15,16]

In total 350 respondents out of 2351 were removed from the sample for either (a) making non-monotonic choices (i.e., choosing a Profile with performance level descriptors lower than the alternative); (b) not responding to the choice task; or (c) ranking all three items incorrectly. The final estimation sample includes 2001 respondents.

IV.B. Beta estimates

Estimation of the utility model (Equations (8) and (9)) is done on a sub-sample of respondents who (a) complied with monotonicity assumptions, (b) responded to the choice task, and (c) ranked at least one of the items correctly in the ranking exercise. The sample includes 2001 of 2351 respondents given a choice task. I begin by showing results for Equation (9), where randomly assigned age descriptors δ_a for ages 9, 13, and 17 are interacted with reading and math performance level descriptors θ_{sl}, providing 24 (3 ages x 4 betas x 2 subjects) point estimates plus a constant. The interaction terms allow us to see whether respondent preferences for performance levels are sensitive to profile age. Results are displayed in Figure IV.

[Insert Figure IV Here]

[15]See Lancsar and Louviere, 2006; Lancsar and Louviere, 2008; Inza and Amaya-Amaya, 2008 for discussion. These papers discuss both monotonicity violations as well as other violations of rational choice theory. The focus is primarily on repeated observation of respondent choice behavior, when preferences should be transitive and consistent. In cases where transitivity and consistency are violated, deletion of respondent choice data is discouraged. Guidelines for best practices in cases of monotonicity violations are not well specified. Higher quality (and more expensive) data can be obtained in order to determine whether respondents failed to comprehend, did not take the choice task seriously, or had other reasons for preferring less over more achievement.

[16]In pilot surveys that took place between September 2014 and June 2015, I attempted to make the performance level descriptors more concise in order to improve respondent comprehension and to present respondents with additional choice sets. This procedure has the drawback of undoing the scale anchoring process described previously. In particular, complete descriptors have already been criticized for excessive unidimensionality, and any additional concision would bolster those criticisms. In an effort to allow for the descriptors to maintain their multidimensionality and increase concision, respondents were randomly assigned sub-elements within each performance level descriptor. I generated 3 to 5 sub-descriptors for each complete performance level descriptor and randomly assigned those. An average estimate of the sub-descriptors would, in theory, describe the multidimensional aspects of full descriptor. Nevertheless, I found that respondents were not additionally successful at ranking sub-descriptors relative to the full performance level descriptor; indeed, for many of the sub-descriptors I constructed, respondents were much worse at ranking them. For these reasons, I chose to use the full descriptor.

The common intercept α anchors point estimates for Math and Reading ages 9, 13 and 17 at Level 150.[17] Because each of the subjects are estimated simultaneously, it is possible to compare across subject domains, as well as within domain, across performance level. The solid and dashed lines correspond to fitted quadratic and cubic regression lines, precision weighted by the inverse of the standard error squared. Analytic weights are likewise applied to each of the point estimates to indicate precision (i.e,. larger circles have smaller standard errors).

With only five estimated points, there are many data missing throughout the entire range of potential scale scores. The problem of missing data is unique to educational settings, where two measures of ability are commonplace: discrete performance level descriptors and continuous measures. In order to capture the full continuous range of ability using only discrete descriptors, we will need to fill in the missing data. The two interpolation and extrapolation schemes presented here (quadratic and cubic interpolation) are seen to be inadequate. Cubic interpolation does not impose monotonicity on the interpolated line, thus violating axioms. Quadratic interpolation does not represent the curvature of the line well. The primary purpose of Figure IV is to illustrate two problems with the schemes. Finally, in figures not shown, using either extrapolation method for points beyond 150 and 350 leads to outlandish prediction.

To correct these limitations, I use monotone piecewise cubic interpolation (MCI) as suggested by Fritsch and Carlson (1980). MCI produces results depicted in Figure V for the range 100 to 500. The top panel shows results for Equation (8), where the age-interaction terms are removed. By construction, the curvature is monotonic throughout the entire range and fits the estimated data perfectly. MCI extrapolates for points beyond 150 and 350 by linearly fitting a line from the last two known points (i.e., 150 to 149 and 349 to 350). Linear extrapolation may not be appropriate for points outside the estimated range. Later, I test how sensitivity results are to alternative extrapolation techniques.

[Insert Figure V Here]

We can now observe results. First note the concavity of the each of the point estimates.

[17] It is not possible to disaggregate α into Reading and Math Levels 150, as respondents are forced to make a math and reading choice simultaneously.

As hypothesized, welfare returns to achievement are non-linear and decrease at the higher end of the scale. This is true for all ages and subjects. There is variation in the curvature between subjects and ages. For all ages in reading (bottom right panel of Figure V), there is a steep gain in utility at the bottom of the scale, and then utility gains flatten out. Age 17 shows a steep increase at the high end of the scale, but much of this is due to extrapolation beyond the estimated value of 350. For math (bottom left panel), the largest gains are in the middle of the distribution, as scores increase from 200 to 350, and this is true for all ages. Overall, we see confirmation of the initial hypothesis that utility gains for achievement are non-linear and concave.

In order to estimate changes in achievement across age, by cohort, it will be necessary to combine age terms and estimate Equation (8). Recall the monotonicity assumption implicit in the model: increases in the original scale score must be associated with increases in benefit. As seen in Figures IV and V, each of the age curves are monotonically increasing, but the model does not impose monotonicity across age. To understand why, consider Figure VI, which shows a stylized depiction of Figure V overlaying Ages 9 and 17. Here, the curvatures for Ages 9 and 17 are respectively monotonic, but as achievement along the x-axis increases and "jumps" from Age 9 to 17, there is a concomitant decrease in Y, i.e. utility. This violates modeling assumptions and implies that as children gain in achievement as they age from 9 to 17 they are made worse, all things considered. This implication is made despite the fact that we observe positive utility returns to achievement *within* age.

[Insert Figure VI Here]

We observe this "jump" problem because respondents are not asked to make marginal welfare preferences for achievement *gains* but are instead asked to state preferences for achievement *states*. The theoretical and practical differences between estimating gains and states is a recurring theme in health research and was introduced earlier. The problem is even more pronounced in educational applications, as any vertical scale assumes change in ability across age. Nevertheless, it is not obvious whether welfare evaluations should be sensitive to those changes. Moreover, most test scores are presented as cross-sectional measures of ability. Given that the purpose of this exercise is to convert a commonly used measure

of ability into one that connotes utility, using the cross-sectional achievement score seems appropriate. Modeling growth may be possible but is left aside for future research. [18]

Point estimates will therefore be taken from Equation (8). By eliminating the age interaction terms, the model describes average welfare returns to achievement and is monotonic. Point estimates correspond to the weighted average of the three age terms (9, 13, 17) for each performance level. This can be seen in the upper left and right panels of Figure V. Comparing between lower and upper panels of Figure V shows that age-specific point estimates do not substantively alter interpretation. Moreover, ignoring the age-interactions helps to mitigate some of the exaggerated extrapolation for ranges beyond 350.

IV.B.1. Estimating and converting NAEP scales

Here I describe how estimated utility values for Reading and Math performance level descriptors 150, 200, 250, 300 and 350 are applied to individual level NAEP data. To do this, I take individual level data from the NAEP restricted-use files and generate a vector of reading and math scores for each individual student's 5 plausible values.[19] Each individual student's score is estimated according to the MCI projection. This is done for all student scores in reading and math, ages 9, 13 and 17, for years 1990-2008. As a summary statistic, I take the mean NAEP and mean welfare-adjusted score for each subject, age, year and subgroup, taking account of the NAEP's complex survey design as well as the five plausible values.[20] Finally, in order to compare the original scale, which ranges between 100 and 500, to the welfare-adjusted scale, which ranges from -2.7 to 0.2, I standardize them both to have mean $\mu = 100$ and standard deviation $\sigma = 10$.

[18] See Weinstein, et al., 2009 and especially Nord, et al., 2009 for important discussion about gains versus levels in health, with emphasis on both policy and normative implications.

[19] In the NAEP, individuals do not receive the complete battery of tests. For this reason, each individual student is given 5 plausible values which are randomly drawn from a distribution of possible θ values. The 5 plausible values can be combined to provide summary statistics for sub-populations following Rubin's rules for multiple imputation. See Mislevy, et al., 1992 for a description of this procedure.

[20] Specifically, to estimate means for each plausible value of the NAEP, I use Stata's –svy– commands, specifying probability and replicate weights, as well as the sampling clusters. I follow Rubin's rules to aggregate across each of the 5 plausible values. The mean score is a simple average of each of the subject-age-year score, but error variance requires that we take account of between plausible value variation and the error variance of each estimate. The formula for this is: $within = \frac{1}{5}\sum_{p=1}^{5}\sigma^2$; $between = \frac{1}{4}\sum_{p=1}^{4}(\bar{X} - X_p)$; $total = \sqrt{within + 1.2 * between}$.

IV.C. Comparing original to welfare-adjusted scale

I now show how inferences between the original NAEP scale and the estimated and interpolated welfare-adjusted scale contrast. I first present a stylized figure to show the kinds of cases for which inferences between the two scales will diverge. I then compare white-black achievement gaps (defined as the mean difference between the two groups) across cohort (that is, as students age).

IV.C.1. Achievement gap example

An example of a change in math achievement for a single cohort is shown in Figure VII. The x-axis shows the original standardized NAEP scale and the y-axis shows the estimated and interpolated scale for a cohort of students in years 1982 to 1990. The solid intersecting lines indicate mean black scores and the dashed intersecting lines indicate mean white scores; the scores at the lower end of the distribution are for students at age 9 and at the higher end of the distribution are for students at age 17. The difference between dashed and solid lines for the respective axes provides the achievement gap.

[Insert Figure VII Here]

It is clear from Figure VII that white black differences at age 17 are slightly smaller than they were at age 9 for both scales, indicating that the gap shrank as children aged. The size of the change in the gap is much smaller using the welfare-adjusted scale than it is using the original scale, as the difference in scores at age 9 are smaller in the welfare scale than they are in the original NAEP. What is also revealing about this figure is that if all student scores increased by the same amount (i.e., an equivalent mean increase in achievement), the effect on the achievement gap in the adjusted scale would be profound. By shifting all scores to the right, the size of the gap at age 9 in the adjusted scale would be larger, as a result of the steeply increasing value in achievement, and the size of the gap at age 17 would be smaller, as a result of the fact that gains at the high end of the scale are diminished. Taken together, the adjusted scale would show a dramatic decrease in achievement gaps between the ages of 9 and 17, simply by increasing all scores an equal amount. This contrast between scales is

exactly the consequence we would expect when equal interval assumptions with respect to welfare are violated.

One other point is worth emphasizing from Figure VII. The first is that differences in inferences between the two scales requires change. Cross-sectional comparisons between the two scales result only in intercept differences (e.g., a score of 80 in the original scale is equivalent to a score of 83). While achievement gaps have narrowed somewhat over time, most of the change in NAEP scores takes place as children age. For that reason, I show gap changes as children age, across cohort.

IV.C.2. Achievement gaps: Across cohort

The effects of rescaling can be seen when we look at achievement gap changes as children age. The NAEP is vertically equated, meaning examinees at ages 9, 13 and 17 are given a overlapping sample of test items at each age level. While there are some concerns about the nature of the inference one can draw from vertical equating, such cross-age comparisons are technically allowable with NAEP data (Haertel, 1991). In order to make cross-age comparisons, I use a sub-sample of cohorts for whom a random sample of students are tested at age 9 in year t and tested again at age 17 in year $t + 8$. The achievement growth for students from age 9 to 17 in year t to $t + 8$ is provided for six cohorts c in both math and reading. Within each of these cohorts, because samples of students are randomly drawn in each interval, it is possible to say that the achievement of any subgroup g in cohort c grew or shrank by some amount, using both the original NAEP and welfare-adjusted scales. The achievement gap for any cohort c is defined as the mean white minus mean black score in years t and $t + 8$.

[Insert Figure VIII Here]

[Insert Figure IX Here]

Figures VIII and IX present results for the possible six cohorts in math and reading, respectively. Solid lines depict the original NAEP scale and dashed lines depict the welfare adjusted scale. As hypothesized, in many instances, the inferences we would draw from the adjusted scale depart in magnitude and sign when compared to the original scale. In math,

the 1982 to 1990 cohort (depicted in green) is as described in Figure VII, and we observe here what was described there: a rate of gap closure that is steeper in the original NAEP metric than in the welfare scale. In 1978-1986, 1992-2000, and 1996-2004 cohorts gap signs are reversed. Whether or not trends are reversed between the two scales will be a function of the size, location and rate of change of the subgroup's respective mean achievement.

In reading, the departure between the original and adjusted scales is much more pronounced. Using the adjusted scale, the reading achievement gap is shown to be decreasing by about 6 to 10 points for every cohort. Conversely, using the original scale, the gap is decreasing by about 1 to 3 points in four cohorts and increasing by 1 to 2 points in two cohorts. This can be traced back to Figure V, where we observed a steep change in slope beginning at NAEP score 250. Gains below 250 are very steep, while gains above are much more shallow. If black scores, between ages 9 and 17, move along the back half of the curve, while white scores moves along the front half of the curve, gap decreases will be much larger.

V. Conclusion

Overall, I have demonstrated that welfare benefits of different achievement states described by the NAEP are not equal interval. In contrast to existing methods, the technique I propose provides a direct and explicit description of the welfare gains from different achievement states. Moreover, instead of linking achievement to earnings, I have suggested that the benefits of achievement can be described inclusively, meaning that achievement need not serve merely pecuniary purposes. With the proposed method, the inferences we draw about changes in achievement and changes in achievement gaps (especially as children age) will differ depending on which scale we use. Which scale ought to be used is, I have argued, application sensitive. When descriptions of academic ability are desired, or when we wish to know how much more or less some subgroups know about math and reading relative to other subgroups, the original NAEP scale can allow for such inferences. When, however, we wish to derive some additional inference about the scale—for instance, when an achievement score is used as an outcome variable, when a score is used for cost-effectiveness evaluation,

or when we wish to evaluate whether a narrowing of the achievement gap is "good" or "bad"—the original NAEP scale is inadequate. It fails to accurately describe benefit in any meaningful way. In this paper, I have described and implemented a method that allows for such values-based inferences.

In light of the previous discussion, we can revisit the four questions that were raised at the start of this essay.

1 To what outcome should scale scores be indexed?

2 Whose preferences for achievement should be included in the index?

3 How should the index balance individual and social benefits?

4 How should time be modeled in the elicitation and estimation of the utility value?

I have supplied answers to each of these questions. Outcomes are indexed to survey respondents' understanding of how much welfare is attributable to certain levels of achievement; college educated respondents are included in the index; equity is given zero weight in the model; time is modeled cross-sectionally. Whether or not these choices are the correct ways to link achievement to outcomes is not known, but the choices inherent to the inference are here made explicit.

Contrast the approach detailed here to when achievement scores are used as outcome variables. With the use of achievement scores, even if the equal-interval assumptions hold, the implicit assumptions of the model are that benefits are best characterized by ability differences, that all ability differences are equally beneficial, and that all benefits are individual (and not societal) and best characterized by a cross-section in time. These assumptions lack theoretical and, as demonstrated here, empirical warrant; nevertheless, these assumptions form the basis of a great majority of education policy evaluations. Education policy evaluation will be greatly improved when the implicit assumptions underlying the use of traditional achievement scores are made explicit.

As stated previously, the approach taken here should be interpreted as proof of concept. Many questions remain, and the assumptions used to construct this scale may not be suitable. It is hoped that the procedure be used as a jumping-off point for future inquiry and research.

References

Ballou, D. (2009). "Test scaling and value-added measurement." *Education*, *4*(4), 351–383.

Beaton, A. E. & Allen, N. L. (1992). "Interpreting scales through scale anchoring." *Journal of Educational and Behavioral Statistics*, *17*(2), 191–204.

Beaton, A. E. & Zwick, R. (1992a). "Overview of the national assessment of educational progress." *Journal of Educational and Behavioral Statistics*, *17*(2), 95–109.

——— (1992b). "Overview of the national assessment of educational progress." *Journal of Educational and Behavioral Statistics*, *17*(2), 95–109.

de Bekker-Grob, E. W., Ryan, M., & Gerard, K. (2012). "Discrete choice experiments in health economics: a review of the literature." *Health economics*, *21*(2), 145–172.

Bond, T. N. & Lang, K. (2013a). "The Black-White education-scaled test-score gap in grades k-7."Technical report, National Bureau of Economic Research.

——— (2013b). "The evolution of the Black-White test score gap in Grades K–3: The fragility of results." *Review of Economics and Statistics*, *95*(5), 1468–1479.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). "Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data?" *Perspectives on psychological science*, *6*(1), 3–5.

Cunha, F. & Heckman, J. J. (2008). "Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation." *Journal of Human Resources*, *43*(4), 738–782.

Cunha, F., Heckman, J. J., & Schennach, S. M. (2010). "Estimating the technology of cognitive and noncognitive skill formation." *Econometrica*, *78*(3), 883–931.

Domingue, B. (2014). "Evaluating the equal-interval hypothesis with test score scales." *Psychometrika*, *79*(1), 1–19.

Drummond, M. F. (2005). *Methods for the economic evaluation of health care programmes.*: Oxford university press.

Fritsch, F. N. & Carlson, R. E. (1980). "Monotone piecewise cubic interpolation." *SIAM Journal on Numerical Analysis*, *17*(2), 238–246.

Haertel, E. H. (1991). "Report on TRP Analyses of Issues Concerning Within-Age versus Cross-Age Scales for the National Assessment of Educational Progress.."

Hainmueller, J. & Hopkins, D. J. (2014). "The hidden american immigration consensus: A conjoint analysis of attitudes toward immigrants." *American Journal of Political Science*.

Inza, F. S. M., Ryan, M., & Amaya-Amaya, M. (2007). ""Irrational" stated preferences." *Using Discrete Choice Experiments to Value Health and Health Care*, *11*, 195.

Kuziemko, I., Norton, M. I., & Saez, E. (2015). "How Elastic Are Preferences for Redistribution? Evidence from Randomized Survey Experiments." *American Economic Review*, *105*(4), 1478–1508.

Lancsar, E. & Louviere, J. (2006). "Deleting 'irrational' responses from discrete choice experiments: a case of investigating or imposing preferences?" *Health economics*, *15*(8), 797–811.

——— (2008). "Conducting discrete choice experiments to inform healthcare decision making." *Pharmacoeconomics*, *26*(8), 661–677.

Lee, J. Y. & Solon, G. (2011). "The fragility of estimated effects of unilateral divorce laws on divorce rates." *The BE Journal of Economic Analysis & Policy*, *11*(1).

Lissitz, R. W. & Bourque, M. L. (1995). "Reporting NAEP results using standards." *Educational Measurement: Issues and Practice*, *14*(2), 14–23.

McFadden, D. (1986). "The choice theory approach to market research." *Marketing science*, *5*(4), 275–297.

——— (2001). "Economic choices." *American Economic Review*, 351–378.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). "Scaling procedures in NAEP." *Journal of Educational and Behavioral Statistics*, *17*(2), 131–154.

Mullis, I. V. & Jenkins, L. B. (1988). *The Science Report Card: Elements of Risk and Recovery. Trends and Achievement Based on the 1986 National Assessment.*.: ERIC.

Nielsen, E. R. (2015). "Achievement Gap Estimates and Deviations from Cardinal Comparability." *Available at SSRN 2597668*.

Nord, E., Daniels, N., & Kamlet, M. (2009). "QALYs: some challenges." *Value in Health*, *12*(s1), S10–S15.

Oppe, M., Devlin, N. J., & SZENDE, A. (2007). *EQ-5D value sets: inventory, comparative review and user guide.*: Springer.

Reardon, S. F., Valentino, R. A., & Shores, K. A. (2012). "Patterns of literacy among US students." *The Future of Children*, *22*(2), 17–37.

Solon, G., Haider, S. J., & Wooldridge, J. M. (2015). "What are we weighting for?" *Journal of Human resources*, *50*(2), 301–316.

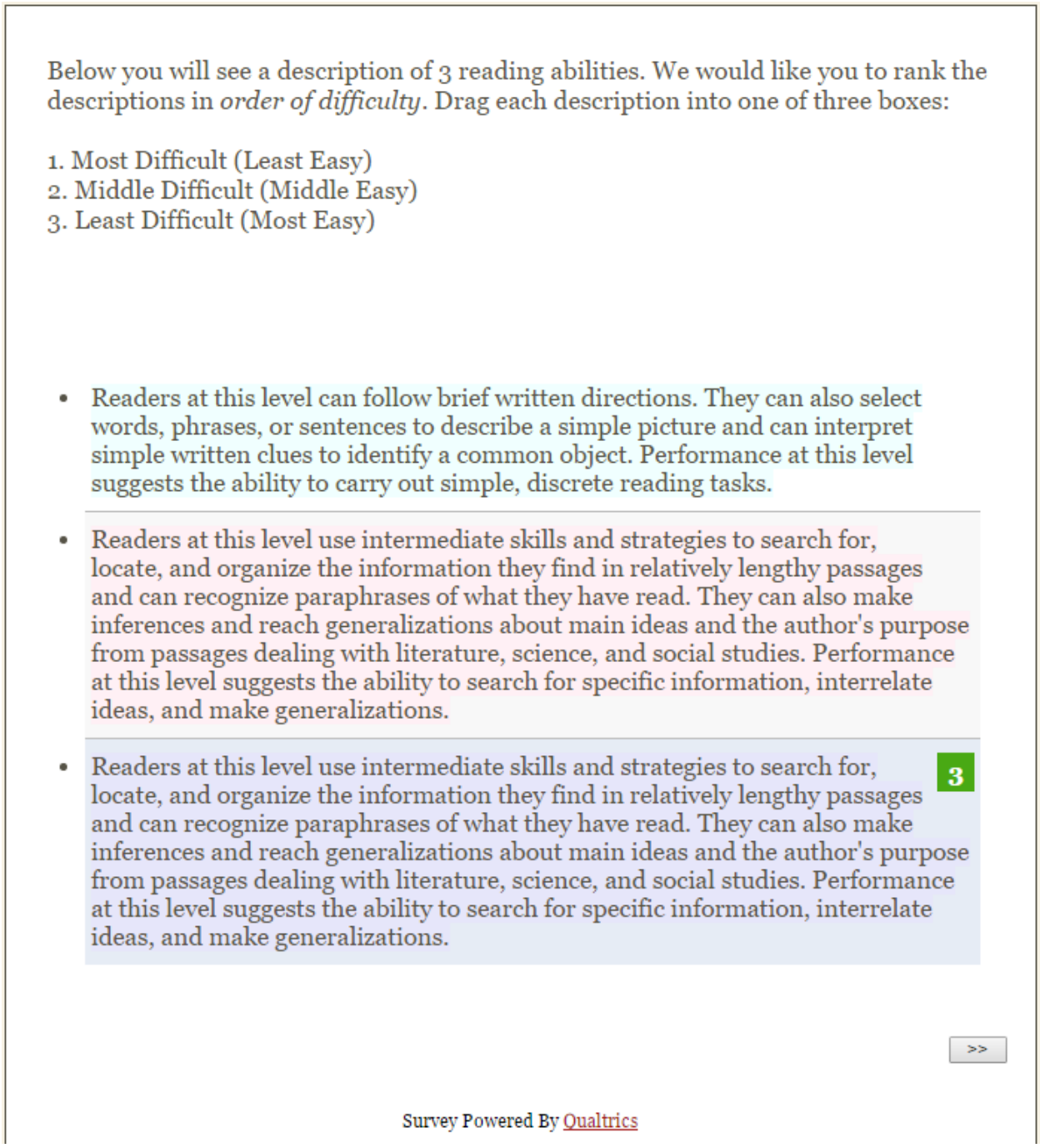
Torrance, G. W., Feeny, D. H., Furlong, W. J., Barr, R. D., Zhang, Y., & Wang, Q. (1996). "Multiattribute utility function for a comprehensive health status classification system: Health Utilities Index Mark 2." *Medical care*, *34*(7), 702–722.

Weinstein, M. C., Torrance, G., & McGuire, A. (2009). "QALYs: the basics." *Value in health*, *12*(s1), S5–S9.

Whitehead, S. J. & Ali, S. (2010). "Health outcomes in economic evaluation: the QALY and utilities." *British medical bulletin*, *96*(1), 5–21.

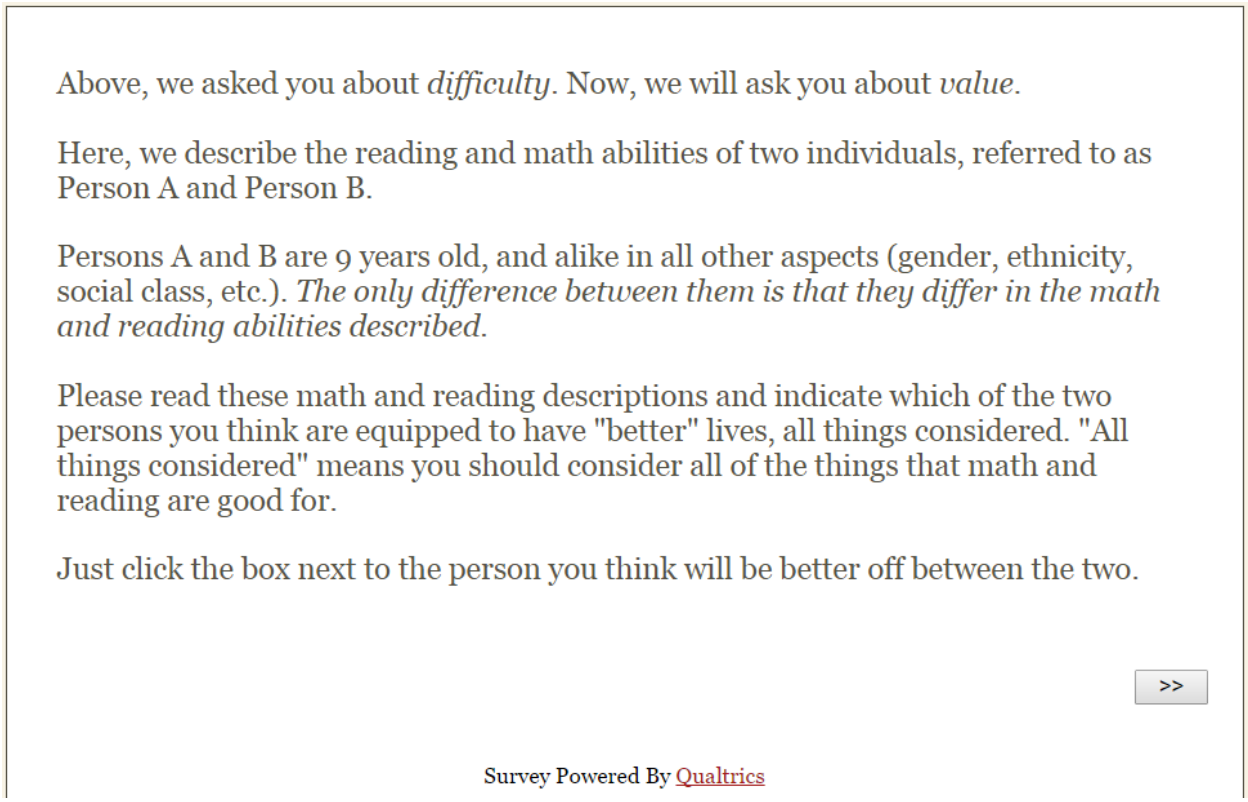
VI. Figures

Figure I: Survey Example: Ranking Exercise



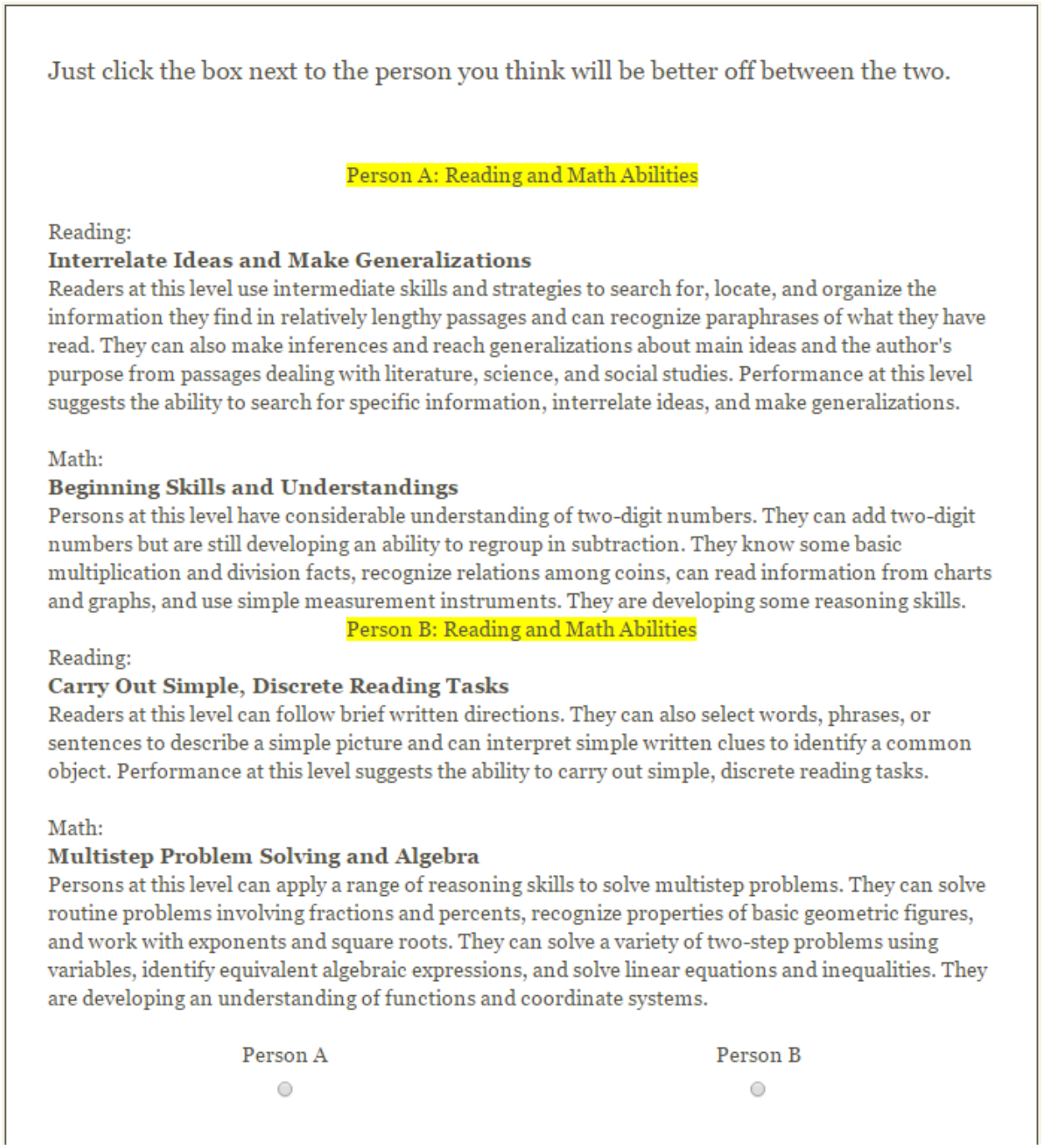
This is a screen shot (1 of 3) from the online survey experiment administered to 2351 respondents through Amazon's Mechanical Turk software. This task asked respondents to rank 3 reading performance level descriptors in terms of difficulty. Respondents were randomly assigned either reading or math subject and 3 of 5 performance level descriptors (with replacement).

Figure II: Survey Example: Introduction to Choice Exercise



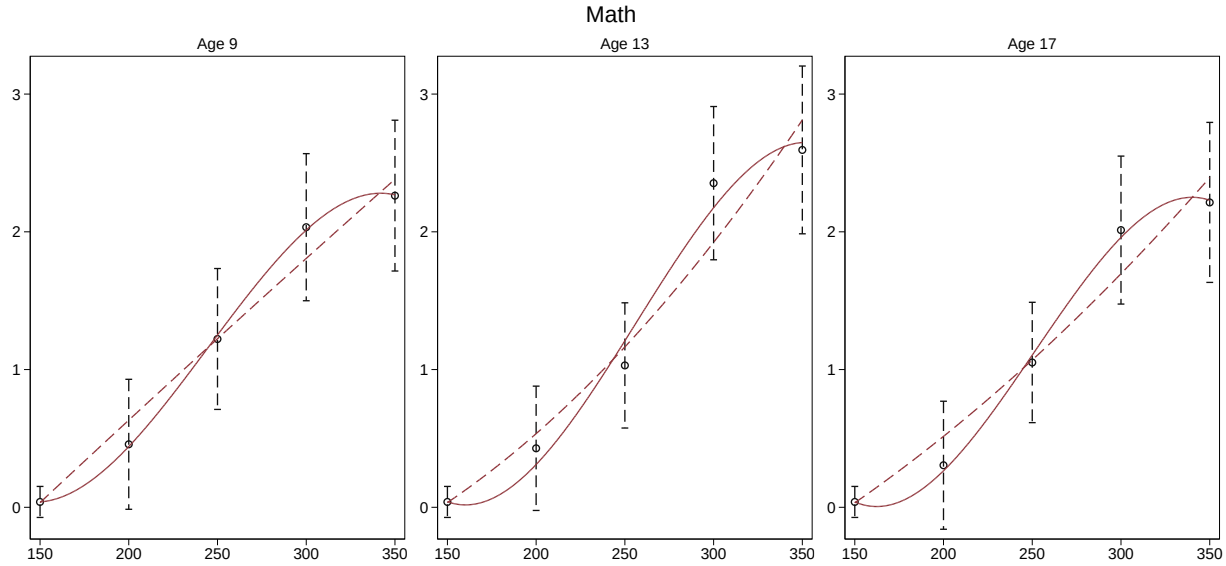
This is a screen shot (2 of 3) from the online survey experiment administered to 2351 respondents through Amazon's Mechanical Turk software. In this screen shot, the choice task is introduced to respondents. Respondents are informed that the two profiles, Persons A and B, are equal in all respects except that they differ in their reading and math abilities. They are instructed to select which person will be better off between the two. In paragraph 3, Persons A and B are also randomly assigned an age, which can be either 9, 13 or 17.

Figure III: Survey Example: Choice Exercise

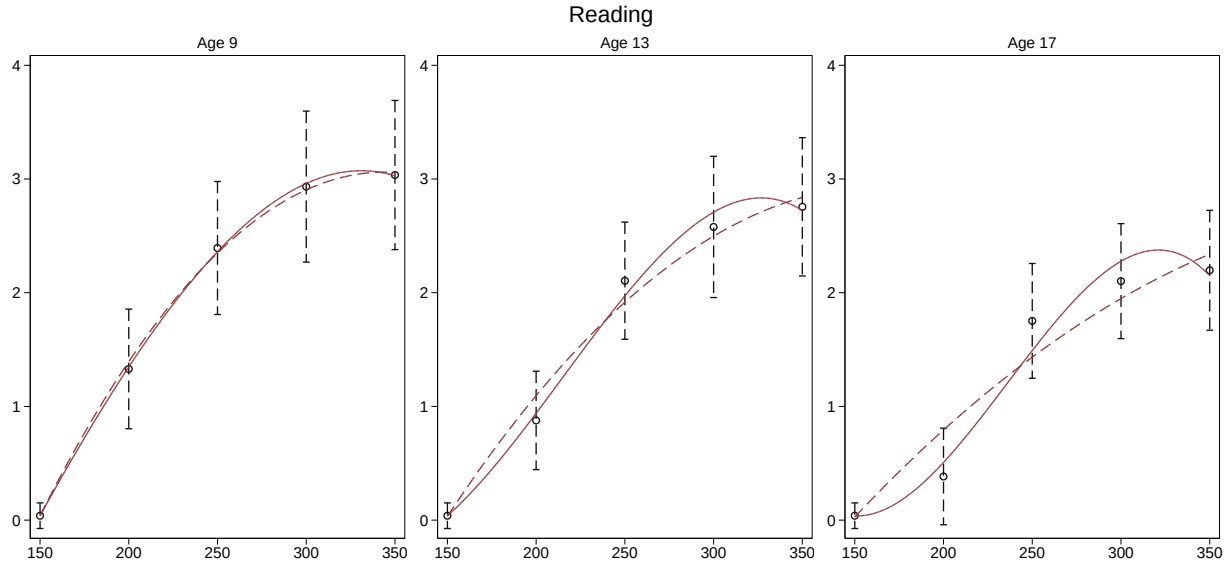


This is a screen shot (3 of 3) from the online survey experiment administered to 2351 respondents through Amazon's Mechanical Turk software. In this screen shot, the choice task is presented to respondents. Respondents are randomly assigned a reading and math performance level descriptor for Persons A and B, with replacement. Performance level descriptors are taken from the NAEP-LTT and can be seen in Tables I and II. At the bottom of the choice task, respondents select which person (A or B) they think would be better off, "all things considered."

Figure IV: Estimated Beta Coefficients for Math and Reading, Age Interactions



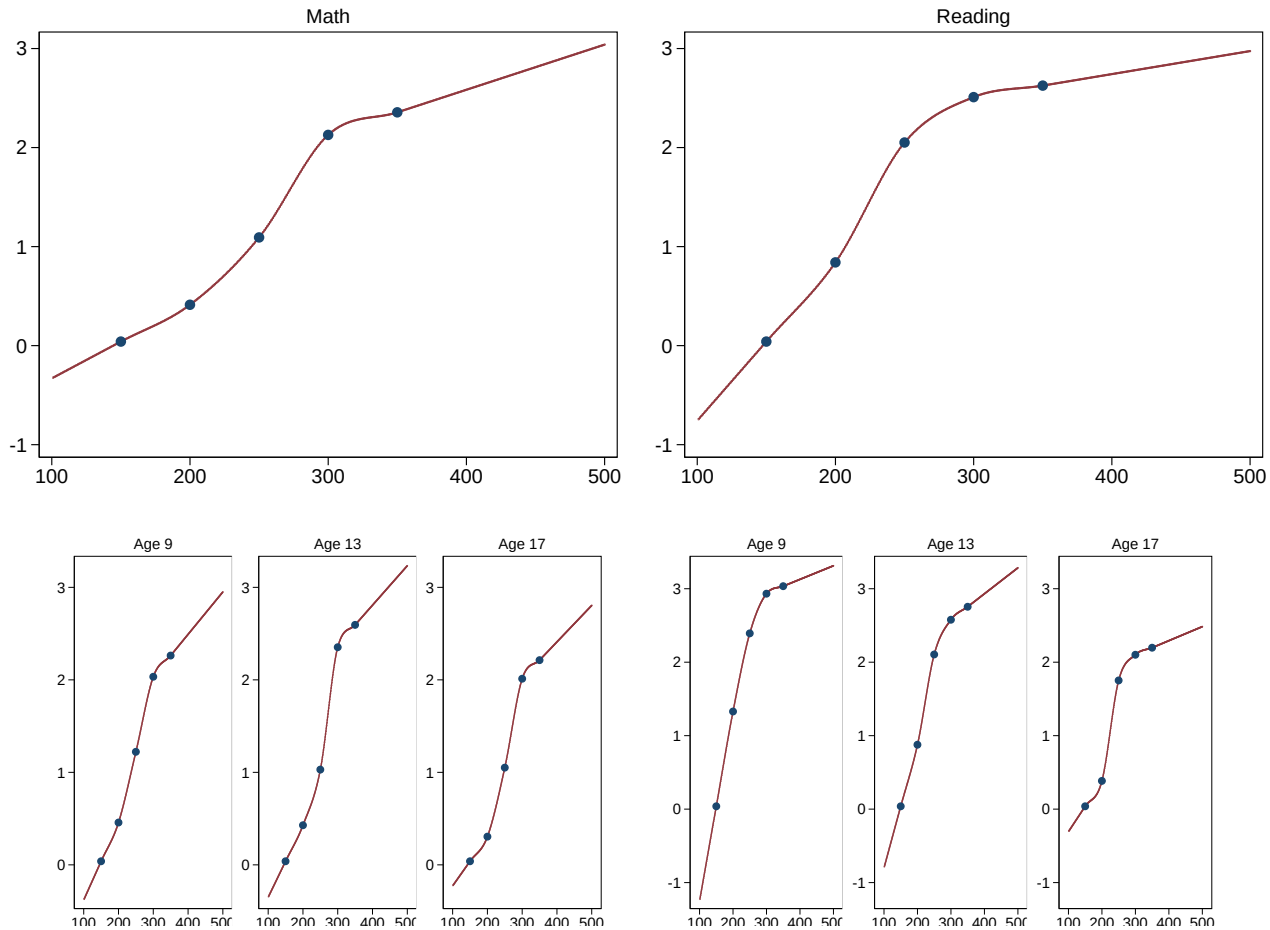
(a)



(b)

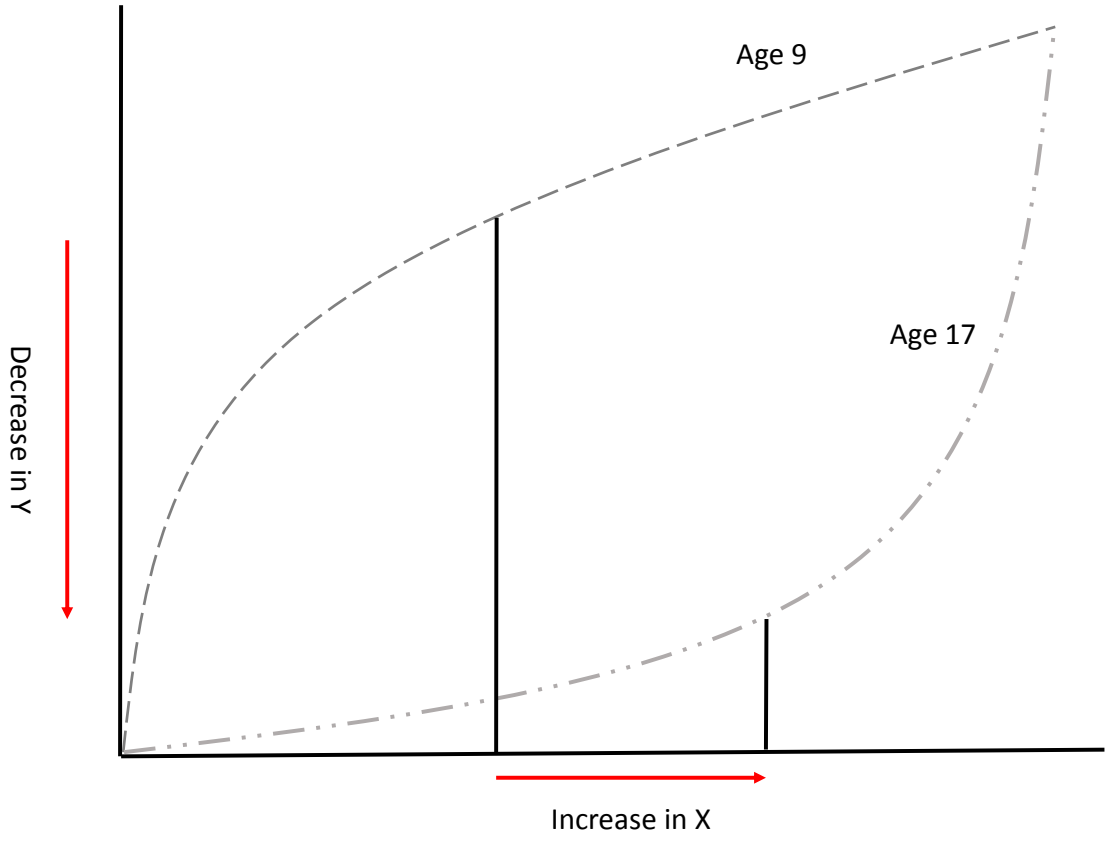
This figure depicts point estimates from logistic regression Equation (9) shown. Point estimates indicate probability of respondent selecting profile with math (top panel) or reading (bottom panel) performance level descriptor equal to 200, 250, 300 or 350 (relative to omitted category 150). Solid line drawn using precision-weighted cubic regression through the estimates; dashed line drawn using precision-weighted quadratic regression. Range gaps depict 95 percent confidence intervals.

Figure V: Monotonic Cubic Interpolation of Math and Reading Beta Coefficients



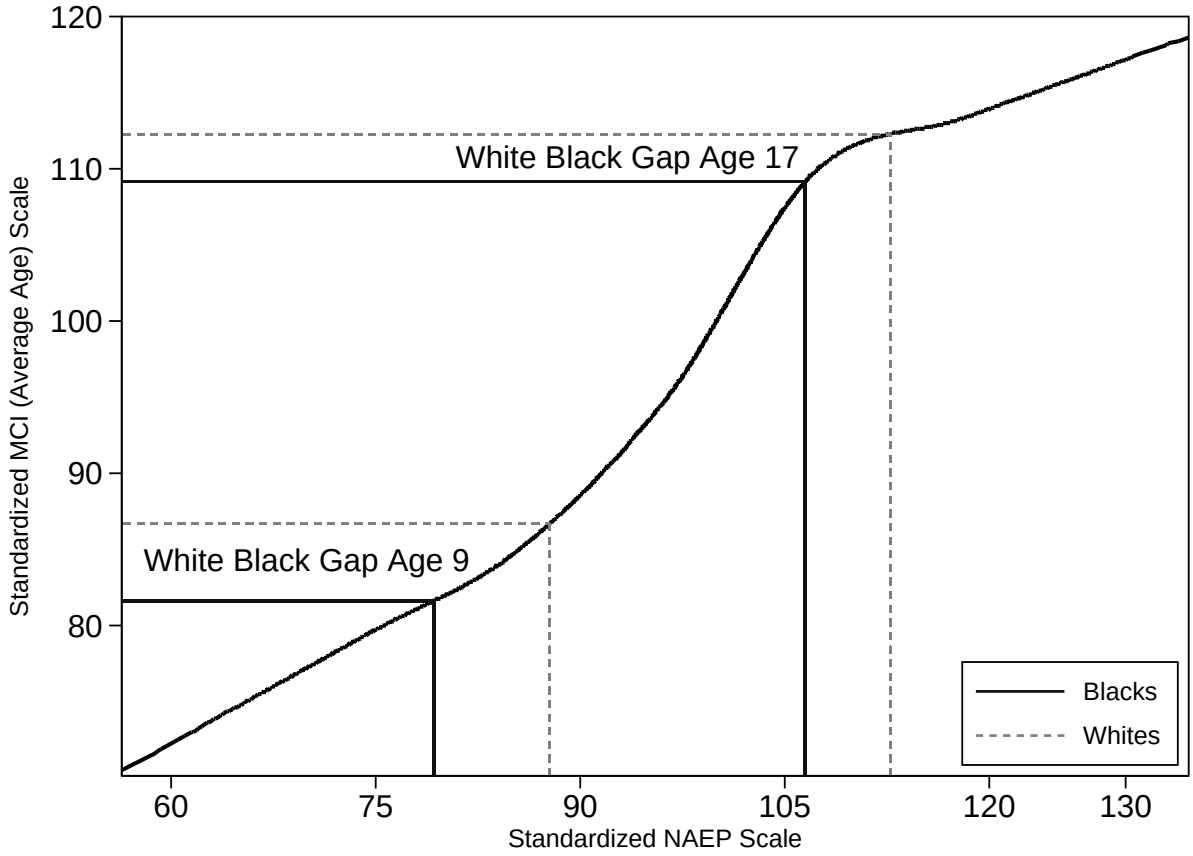
This figure takes point estimates from Equations 8 and 9 and performs piecewise monotone cubic interpolation (MCI) according to Fritsch and Carlson (1980) for scale range 100 to 500. Extrapolation for points less than 150 and greater than 350, respectively, is done via linear extrapolation of the two most proximal points, e.g. linear extrapolation based on points 151 and 150 and 349 and 350, respectively. Top panel drops age interactions (Equation 8) and bottom panel estimates equation 9.

Figure VI: Changes in Scale and Welfare Scores across Age: "Jump" Problem



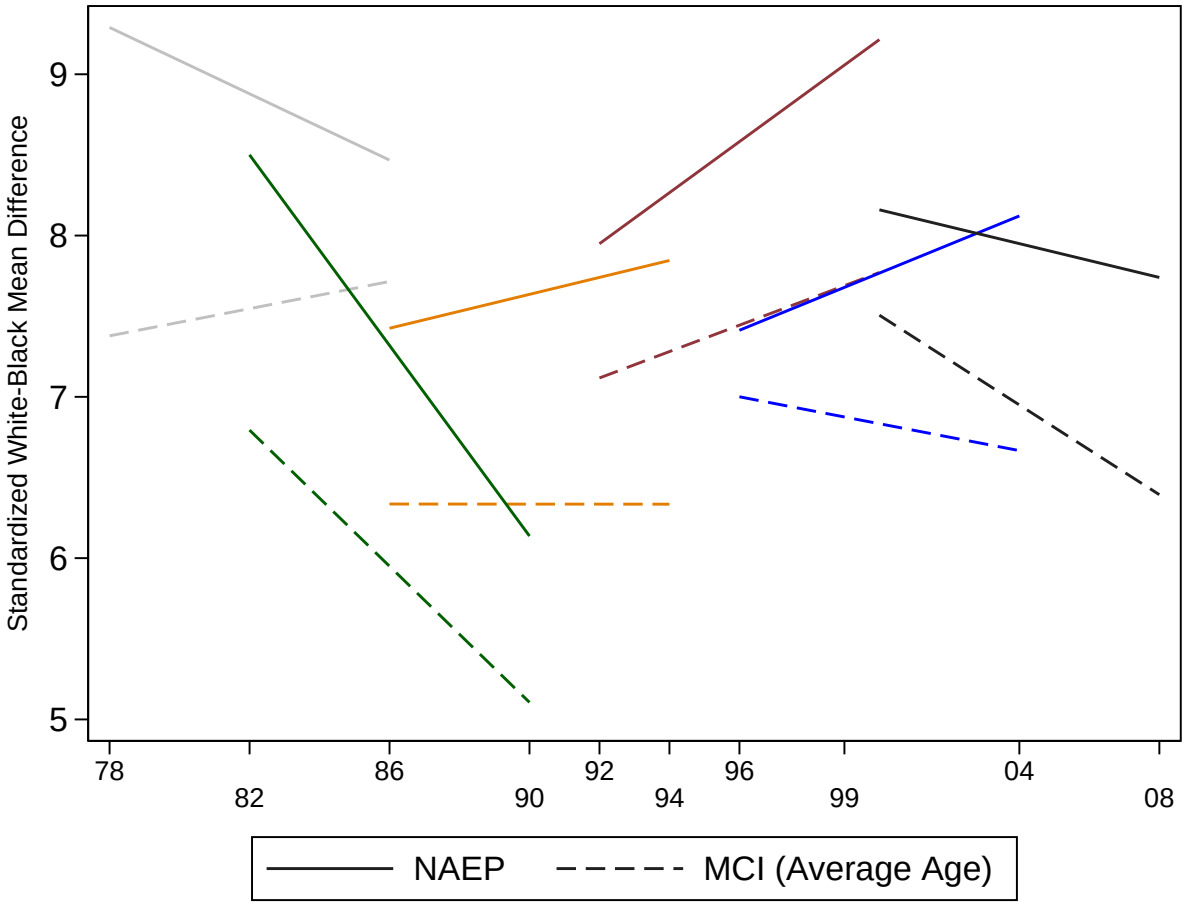
Stylized depiction showing how monotonicity within age need not lead to monotonicity across age, i.e. the "jump" problem. In this representation, benefits are monotonically increasing for ages 9 and 17, but as achievement increases from age 9 to 17, there is a downward "jump" in welfare. This is due to the fact that the choice task is cross-sectional, asking respondents about their preferences for achievement *states* and not achievement *growth*.

Figure VII: White and Black Changes in Math Achievement across Age: Example Cohort



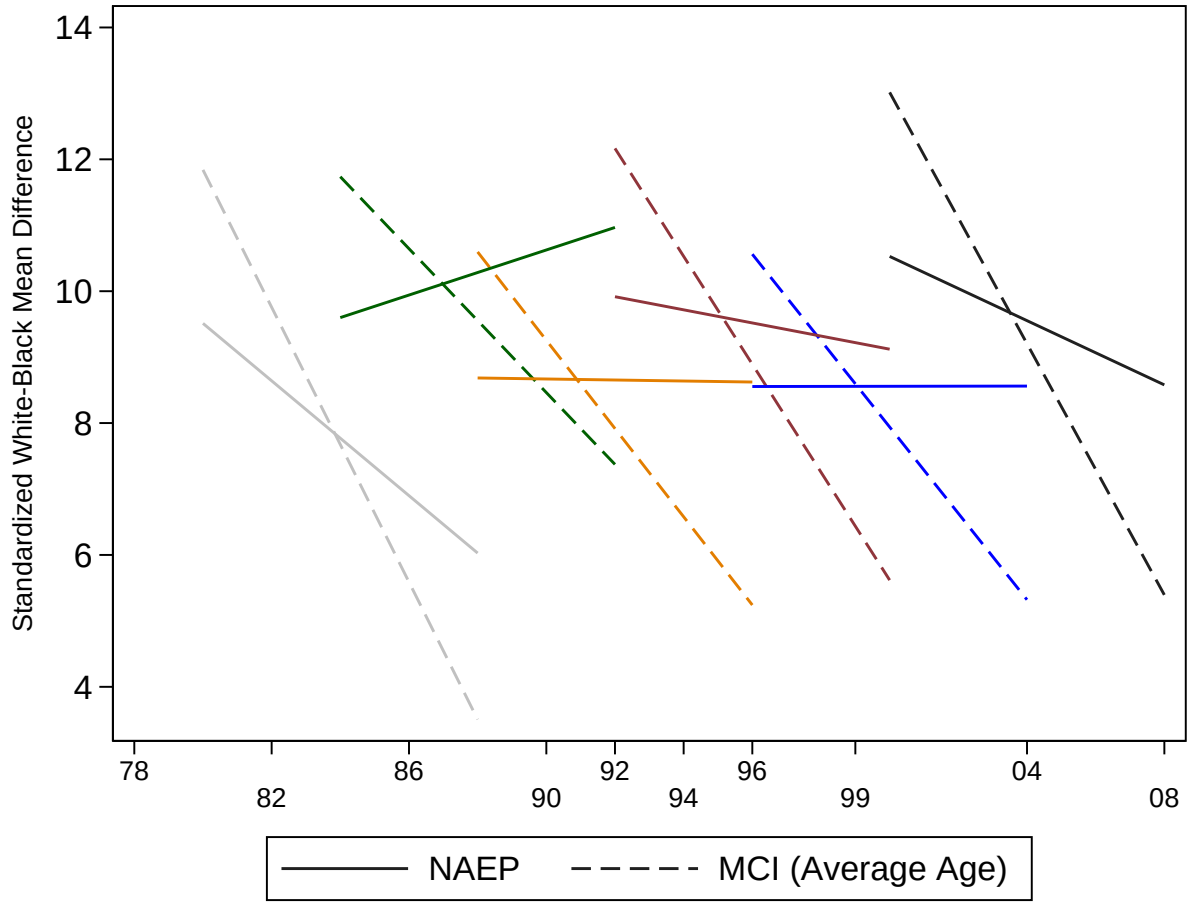
This figure depicts standardized original and welfare-adjusted NAEP scores for one cohort of students ages 9 and 17 for years 1982 and 1990 (for 1 of 5 plausible values). Solid intersecting lines correspond to mean black scores of 9 and 17 year olds in 1982 and 1990, respectively. Dashed intersecting lines correspond to mean white scores for same ages and years. Achievement gaps are represented as the difference between dashed and solid lines at ages 9 and 17 along both the x- and y-dimensions of the graph.

Figure VIII: Mean White Minus Mean Black Math Scores across Age, by Cohort



This figure depicts mean white minus mean black math achievement for six cohorts of students aged 9 and 17 in years t and $t+8$. Solid lines correspond to original NAEP scale; dashed lines to welfare-adjusted scale. Each line reflects change in white-black achievement gap as one cohort of students changes in achievement between the ages of 9 and 17.

Figure IX: Mean White Minus Mean Black Reading Scores across Age, by Cohort



This figure depicts mean white minus mean black reading achievement for six cohorts of students aged 9 and 17 in years t and $t+8$. Solid lines correspond to original NAEP scale; dashed lines to welfare-adjusted scale. Each line reflects change in white-black achievement gap as one cohort of students changes in achievement between the ages of 9 and 17.

VII. Tables

Table I: Reading Performance Level Descriptors

Level 150: Carry Out Simple, Discrete Reading Tasks
Readers at this level can follow brief written directions. They can also select words, phrases, or sentences to describe a simple picture and can interpret simple written clues to identify a common object. Performance at this level suggests the ability to carry out simple, discrete reading tasks.

Level 200: Demonstrate Partially Developed Skills and Understanding
Readers at this level can locate and identify facts from simple informational paragraphs, stories, and news articles. In addition, they can combine ideas and make inferences based on short, uncomplicated passages. Performance at this level suggests the ability to understand specific or sequentially related information.

Level 250: Interrelate Ideas and Make Generalizations
Readers at this level use intermediate skills and strategies to search for, locate, and organize the information they find in relatively lengthy passages and can recognize paraphrases of what they have read. They can also make inferences and reach generalizations about main ideas and the author's purpose from passages dealing with literature, science, and social studies. Performance at this level suggests the ability to search for specific information, interrelate ideas, and make generalizations.

Level 300: Understand Complicated Information
Readers at this level can understand complicated literary and informational passages, including material about topics they study at school. They can also analyze and integrate less familiar material about topics they study at school as well as provide reactions to and explanations of the text as a whole. Performance at this level suggests the ability to find, understand, summarize, and explain relatively complicated information.

Level 350: Learn from Specialized Reading Materials
Readers at this level can extend and restructure the ideas presented in specialized and complex texts. Examples include scientific materials, literary essays, and historical documents. Readers are also able to understand the links between ideas, even when those links are not explicitly stated, and to make appropriate generalizations. Performance at this level suggests the ability to synthesize and learn from specialized reading materials.

Reading Performance Level Descriptors for National Assessment of Educational Progress, Long Term Trend. Available here: https://nces.ed.gov/nationsreportcard/ltt/reading-descriptions.aspx

Table II: Math Performance Level Descriptors

Level 150: Simple Arithmetic Facts
Students at this level know some basic addition and subtraction facts, and most can add two-digit numbers without regrouping. They recognize simple situations in which addition and subtraction apply. They also are developing rudimentary classification skills.

Level 200: Beginning Skills and Understandings
Students at this level have considerable understanding of two-digit numbers. They can add two-digit numbers but are still developing an ability to regroup in subtraction. They know some basic multiplication and division facts, recognize relations among coins, can read information from charts and graphs, and use simple measurement instruments. They are developing some reasoning skills.

Level 250: Numerical Operations and Beginning Problem Solving
Students at this level have an initial understanding of the four basic operations. They are able to apply whole number addition and subtraction skills to one-step word problems and money situations. In multiplication, they can find the product of a two-digit and a one-digit number. They can also compare information from graphs and charts, and are developing an ability to analyze simple logical relations.

Level 300: Moderately Complex Procedures and Reasoning
Students at this level are developing an understanding of number systems. They can compute with decimals, simple fractions, and commonly encountered percents. They can identify geometric figures, measure lengths and angles, and calculate areas of rectangles. These students are also able to interpret simple inequalities, evaluate formulas, and solve simple linear equations. They can find averages, make decisions based on information drawn from graphs, and use logical reasoning to solve problems. They are developing the skills to operate with signed numbers, exponents, and square roots.

Level 350: Multistep Problem Solving and Algebra
Students at this level can apply a range of reasoning skills to solve multistep problems. They can solve routine problems involving fractions and percents, recognize properties of basic geometric figures, and work with exponents and square roots. They can solve a variety of two-step problems using variables, identify equivalent algebraic expressions, and solve linear equations and inequalities. They are developing an understanding of functions and coordinate systems.

Math Performance Level Descriptors for National Assessment of Educational Progress, Long Term Trend. Available here: https://nces.ed.gov/nationsreportcard/ltt/math-descriptions.aspx

Table III: Results from Ranking Exercise

		Reading				Math			
		Mean		Mean-by-Distance		Mean		Mean-by-Distance	
Level 150		0.751	***	0.647	***	0.691	***	0.531	***
		(0.026)		(0.063)		(0.027)		(0.066)	
	Distance 150			0.161	*			0.143	
	Distance 200			0.1				0.228	**
Level 200		0.788	***	0.733	***	0.661	***	0.585	***
		(0.027)		(0.044)		(0.027)		(0.047)	
	Distance 150			0.059				0.062	
	Distance 200			0.153	*			0.258	**
Level 250		0.672	***	0.603	***	0.721	***	0.664	***
		(0.026)		(0.037)		(0.027)		(0.038)	
	Distance 150			0.163	**			0.062	
	Distance 200			0.106				0.238	**
Level 300		0.697	***	0.59	***	0.711	***	0.713	***
		(0.026)		(0.045)		(0.027)		(0.047)	
	Distance 150			0.155	**			0.072	
	Distance 200			0.184	*			-0.223	**
Level 350		0.627	***	0.478	***	0.637	***	0.531	***
		(0.027)		(0.066)		(0.027)		(0.066)	
	Distance 150			0.14				0.136	
	Distance 200			0.197	**			0.122	
N				1455				1461	
Respondents				485				487	

Regression model estimates linear probability that respondents ranked performance level descriptor correctly. Samples excludes respondents if (a) they were randomly assigned "ties" or (b) they did not rank all three items. Column "Mean" describes percent of reading or math level descriptors $\in (150, 200, 250, 300, 350)$ ranked correctly. "Mean-by-Distance" disaggregates percentages into three categories: whether the cumulative distance of the three descriptors summed to 100, 150, or 200 (e.g., a random draw of 150, 200, 250 sums to 100). Stars indicate * for p<.05, ** for p<.01, and *** for p<.001. Mean-by-Distance test is relative to omitted category, Distance 100.