Journal of Methods and Measurement in the Social Sciences,
Vol.14, No. 1, 1-30, 2023

Large Field Surveys and Short Items Sets: Estimating Reasoning Skills among Distinct Samples in Malawi and Zambia

Hannah Silverstein Sudhanshu Handa

David Thissen

University of North Carolina, Chapel Hill

Raven's Progressive Matrices measure logical reasoning and are often included in large multi-topic surveys in low and middle-income countries. The matrices are image-based items that do not require formal knowledge of language or math to complete. As such, they are attractive items to measure logical reasoning in international development contexts. Many of these large field surveys include short item Raven's sets because space is too limited to fit a full suite. However, short sets can result in restricted variation in terms of test scores. In this paper, we use a nominal response model (NRM) form of item response theory (IRT) to uncover hidden variation in right and wrong answers using short-item Raven's tests from two large field surveys in Malawi and Zambia. We also analyze relationships between a set of other variables, comparing performance of different versions of the logical reasoning scores as both independent and dependent variables, checking the validity of the new scores. The new NRM-estimated logical reasoning scores follow a more normal distribution in both samples. Validity checks suggest that when relationships are less strong, NRM-estimated scores can capture more nuance than summed scores or even 2 parameter logistic IRT-estimated scores. NRM can uncover differences that are not apparent when using simple summed scores.

Direct measurement of cognitive and non-cognitive skills in large population-based multi-topic surveys is a challenge. Accurate assessment of these domains typically involves administering time consuming tests or activities which usually cannot be accommodated in a multi-topic survey such as the Demographic and Health Surveys, Multiple Indicator Cluster Surveys, or Living Conditions Survey. Yet the potential benefit of measuring cognitive and reasoning skills in multi-topic surveys is that these measures can then be linked to actual behavioral outcomes such as labor market performance, wages, migration, family formation, and health, thus greatly expanding the research potential of these surveys. In this article, we present an approach for uncovering latent logical reasoning skills from a short item test of select Raven's Progressive Matrices inserted into large sample, multi-topic household surveys in Malawi and Zambia. Our approach uses information contained in incorrect responses. We develop a more nuanced measure of the latent trait which displays greater variation than simple summed scores.

Cognitive capital is a key source of economic growth that can reduce inequalities and deprivation (UNICEF, 2016). Household surveys are "building blocks of rigorous and transparent monitoring" of progress

toward reducing such inequalities, as the large sample sizes allow for a "high level of measurement precision" across a wide variety of topics (Alkire & Samman, 2014). Many large scale, multi-topic surveys contain items from psychometric tests and scales, an example of which are Raven's Progressive Matrices (RPMs). RPMs are nonverbal tests of inductive reasoning based on figural stimuli, intended to measure a subset of non-academic cognitive skills, more appropriately termed logical reasoning (J. C. Raven et al., 1986, 1992). Researchers often explicitly refer to RPMs as measuring cognition (Akresh et al., 2013; Beaman & Magruder, 2012; Charness et al., 2018; De Groot et al., 2015; Dramé & Ferguson, 2019; Dupas & Robinson, 2013; Hicks et al., 2017; Mani et al., 2013; Teivaanmäki et al., 2017; Tuan Pham Thi Lan, 2003; Vogl, 2014). The basic RPM test format involves incomplete images or pictorial patterns with multiple choice options to complete the missing piece (J. Raven, 2008). RPMs have been included in many renowned large scale surveys such as the Indonesia Family Life Survey (Strauss et al., 2004, 2009, 2016); The World Bank Service Delivery Indicators (Martin & Pimhidzai, 2013; Molina & Martin, 2015; Pimhidzai & Martin, 2015; Wane & Martin, 2016; Wane & Rakotoarivony, 2017); the Young Lives surveys (Azubuike et al., 2016); the Mexican Family Life Survey (Rubalcava & Teruel, 2006, 2008, 2013); and HEalth, Ageing and Retirement Transitions study in Sweden (Lindwall et al., 2017).

In these large, multi-topic surveys it is often not feasible to include full RPM sets (containing 36-60 items, depending on the set version). Instead, select Raven's items are most commonly integrated as shortened sets (Beaman & Magruder, 2012; Charness et al., 2018; Dupas, 2011; Hanaki et al., 2016; Mani et al., 2013; Tuan Pham Thi Lan, 2003; Vogl, 2014). However, the practice of including short sets poses several challenges. These short sets often contain too few items for reliable measurement, and items cannot gradually progress in difficulty. For the Raven's test specifically, a smooth progression in item difficulty provides subjects the opportunity to learn the underlying logic within the test (J. Raven, 2008). Abrupt increases in item difficulty could disrupt the learning process. Similarly, short sets containing disproportionate amounts of difficult or easy items have implications for capturing variation. Another challenge is that data collection using short sets in multi-topic surveys likely differs from optimal psychometric research practices. Under such circumstances Raven's items are often administered in uncontrolled and inconsistent settings, such as at home. And finally, latent trait scores are frequently calculated by simply summing (or averaging) the number of items each individual correctly responds too, which does not consider underlying response patterns (McNeish & Wolf, 2019).

Despite these challenges, a strength of many population-based datasets is their large sample size, possibly allowing for latent trait estimation due to the greater opportunity for more nuanced variation. If any number of the

aforementioned challenges presented in a small sample, which would be more typical in psychometric research, latent trait estimation would be impossible and unreliable.

In this article we estimate latent logical reasoning scores by uncovering variation in responses using both right and wrong answer choices to eight Raven's items administered to 2,514 youth aged 13-19 years in a field survey in Malawi. We then replicate this process with a sample from Zambia consisting of 1,180 adult caregivers of young children, achieving similar results. We compare our estimate of latent logical reasoning scores to simple summed scores and find we are able to uncover more nuanced variation for both samples. These results—consistent across two very different populations in terms of age—suggest that researchers can utilize such approaches to estimate scores with greater variation in other settings when short-item Raven's sets are incorporated into multi-topic field surveys.

Methods

The Malawi Study

Data and sample description

For this analysis we use baseline data from the Transfer Project's impact evaluation of the Government of Malawi's Social Cash Transfer Program. The evaluation was a cluster-randomized study which included a total of 3,531 eligible households. Baseline data were collected in June 2013, which included an additional 821 non-eligible households; more information on the program, study design, and sampling can be found in the baseline report (Abdoulayi et al., 2014).

The household survey included a youth module containing eight RPMs. A maximum of three youth per household, ages 13 to 19, directly responded to questions (The Transfer Project, 2013a). The first four items require closure or completion of increasingly complex visual patterns, while the second block of four items requires application of rules that additively combine shapes across rows and columns. A participant was given an example Raven's item and the correct answer. Then the participants completed the eight matrices, verbally providing the data collector their responses. Data collectors were instructed to limit the overall 'test' to 5 minutes, and to allow approximately 30 seconds for the respondent to answer each problem. Since these Raven's items were part of the household survey, they were administered in the participant's residential setting (The Transfer Project, 2013a).

Although the larger randomized study did not ultimately include the ineligible households, we use youth living in both eligible and ineligible

households to estimate reasoning scores in order to maximize the sample size and potential variation (Abdoulayi et al., 2014). Sample characteristics are in Table 1. There are 2,513 youth with complete Raven's responses in the latent estimation segment of our analysis. The average age of the final analysis sample is 15.37 years, 48.71 percent are female (N=1,212), and the mean years of schooling is 4.41.

Table 1
Malawi youth sample characteristics (N=2,513)

	Mean	Standard deviation
Individual		
Female (1=yes)	0.49	0.50
Age (years)	15.37	1.84
Years of schooling	4.41	2.46
Grade level attainment		
No primary	0.06	0.24
Grade 1	0.05	0.22
Grade 2	0.11	0.31
Grade 3	0.16	0.37
Grade 4	0.15	0.36
Grade 5	0.15	0.36
Grade 6	0.12	0.33
Grade 7	0.07	0.26
Grade 8	0.07	0.26
Grade 9	0.02	0.14
Grade 10	0.02	0.13
Grade 11	0.01	0.08
Completed secondary [a]	0.00	0.07
Household		
Member(s) with disability	0.16	0.37
HH expenditures per capita (MWK)	38,915.13	24,948.41
Female main respondent (1=yes)	0.82	0.38
Main respondent literate (1=yes)	0.27	0.44
Main respondent ever attended school (1=yes)	0.40	0.49
Main respondent age (years)	53.08	18.02

Note: [a] 12 youth completed secondary school, equal to 0.48% of the Malawi sample; This table reports the mean value for each descriptive variable for the Malawi sample.

The challenge of summed scores

As mentioned, summed scores are commonly used for estimating skill levels in tests with selected Ravens items (Beaman & Magruder, 2012; Charness et al., 2018; Hanaki et al., 2016; Mani et al., 2013; Tuan Pham Thi

Lan, 2003; Vogl, 2014). Therefore, we first examine the sample distribution of summed scores, meaning the total number each youth answered correctly.

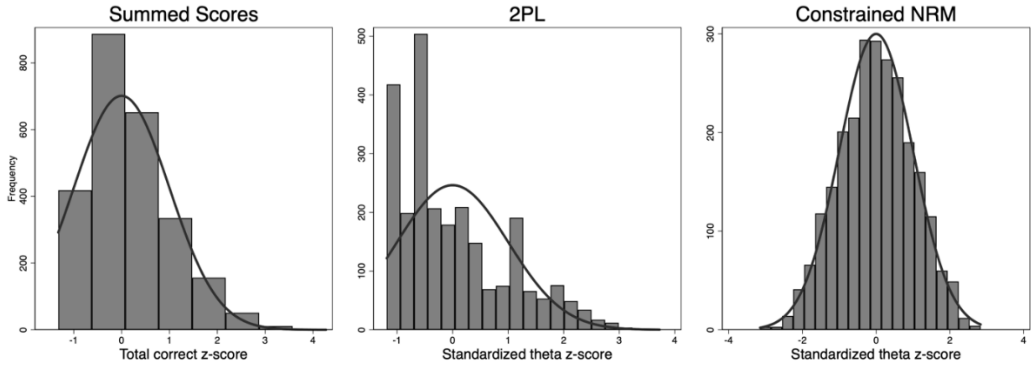
Table 2
Distribution of summed scores among Malawi youth (N=2,513)

Summed score	N	%
0	416	16.73
1	871	35.02
2	650	26.14
3	330	13.27
4	155	6.23
5	51	2.05
6	11	0.44
7	3	0.12
8	0	0.00

Note: This table reports the number and percentage of Malawi youth across all possible summed score values, calculated by totaling the number of correct responses for each participant.

The sample distribution of the summed scores in Table 2 and the left panel of Figure 1 indicate that over half of the sample answered one or zero items correctly. None of the participants answered all eight items correctly. Thus, scores are severely skewed ($p<0.001$), resulting in several implications for using as-is summed scores. Primarily, the skewed distribution indicates a lack of variation, which would make detecting significant results difficult for higher performing participants.

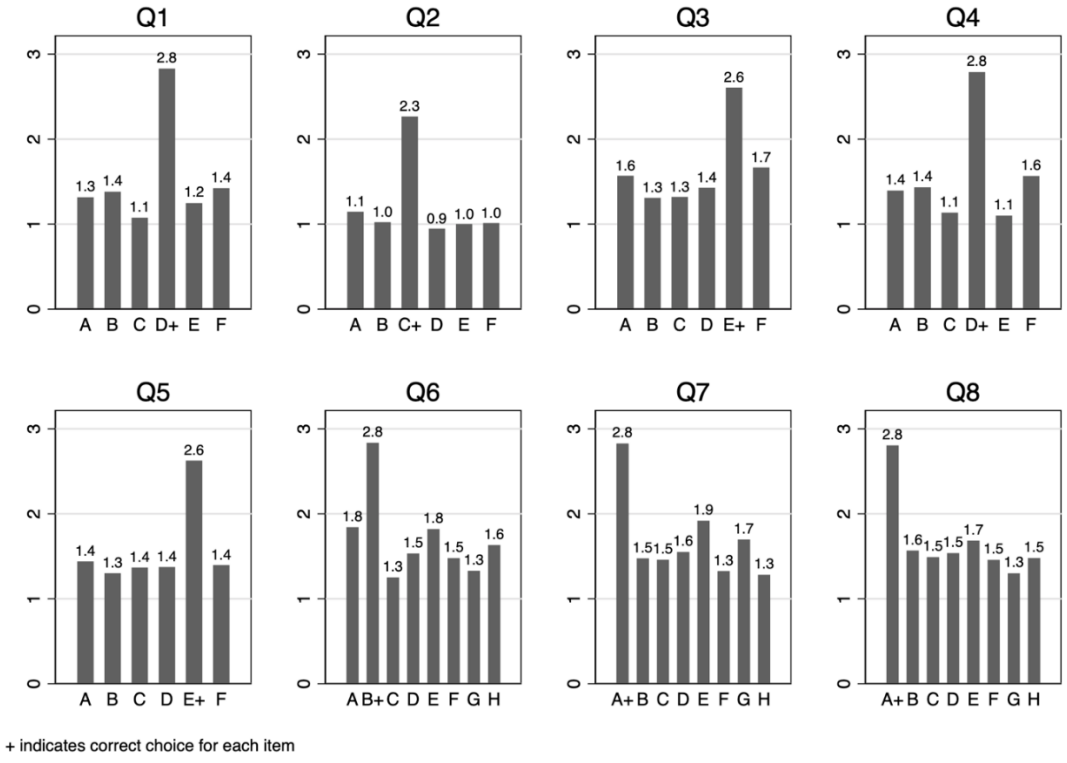
Figure 1. Comparison of standardized distributions of logical reasoning among Malawi youth.



To further highlight this problem, Figure 2 shows the mean sum score for each item option chosen. For all items, the correct choice has the highest

mean summed score. Overall, there is little variation within each question in the average summed score across all incorrect multiple-choice options. This lack of variation forecasts challenges associated with using summed score values as the measure for reasoning skills to predict outcomes in potential multivariate regression models.

Figure 2. Mean summed score for each item response chosen among Malawi youth.



Secondly, since so few participants answered at least five questions correctly, the summed score formulation of logical reasoning is more of a categorical variable than continuous one. Treating the summed score as a continuous variable assumes that each item contributes the same amount of information to the latent trait and that a point increase in the summed score reflects a proportionally linear increase in reasoning skills. This is a common criticism of using summed scores as measures of latent variables (McNeish & Wolf, 2019). The change in format and intended increased difficulty for Questions 6-8 unignorably threatens this assumption: among participants with the same summed score, logical reasoning scores for those who answered these difficult questions correctly should differ from those who answered them incorrectly.

For these reasons, the summed score formulation of reasoning skills is not appropriate in these data and we turn towards more sophisticated approaches used for estimating values of latent variables. Our immediate goals are to reduce the score distribution skew, achieve more nuanced variation in score values that allow for interpretation when using in regression models, and account for varying degrees of difficulty and changes in format between items in the estimation of scores.

Estimating Latent Scores using Item Response Theory (IRT)

Item response theory (IRT) is a mathematical and statistical approach used for testing data to analyze individual items and estimate overall latent trait scores (Thissen & Steinberg, 2009). IRT allows for non-uniformity across test items, which is beneficial in this scenario, given that the last three items had more choices and are intended to be more difficult (Steinberg & Thissen, 2013). IRT analyzes response patterns to discriminate between *higher* and *lower* performing respondents, which is inherently a purpose of scoring. Difficulty and discrimination are the two main parameters estimated in the analyses from this point forward in this paper. θ is the estimated latent variable in IRT, which is logical reasoning in our analysis. Individual scores are measured in standard deviation units of θ (Steinberg & Thissen, 2013).

Our first step estimates a two parameter logistic (2PL) IRT model to see if it improves the distribution of the latent score under a dichotomous variable construction using IRTPRO software (Cai et al., 2011). In this model, the correct answer equals 1 and all other responses for each item are zero-coded. The middle panel in Figure 1 shows the new distribution of latent scores using the 2PL model, which is more nuanced than the sum score distribution, but still heavily skewed ($p<0.001$). However, in 2PL IRT, those only choosing incorrect Raven's responses have the same response pattern, and thus, the same score.

By only considering responses to RPMs as right or wrong, summed scores and dichotomous IRT models ignore potentially useful information captured in the multiple response choices that are incorrect, also known as distractors. Because of the skew and low marginal reliability in the 2PL model, we want to utilize both the correct and distractor choices to analyze response patterns and estimate the latent variable scores. We hypothesize some distractors appeal more to people with higher logical reasoning skills, some distractors are more attractive to people with lower logical reasoning, and some distractors may not differentiate between lower and higher performing respondents. In other words, among the wrong answer choices, some are better than others. And thus, the main question we address in this analysis is whether information about the patterns of wrong answers help

smooth out the distribution and capture more meaningful and nuanced variation.

There is no predetermined order of the distractors for our test and sample. That is, for most of our puzzles, the second-best response after the correct response is not immediately obvious. Therefore, our first challenge is to determine the ordering for the incorrect choices. To minimize the researcher bias imposed on the ordering and score estimation process, we opt for a data-driven approach.

The nominal response model (NRM) does not impose ordering assumptions and is categorical approach to IRT. NRM IRT can be used in two ways: first to empirically determine the ordering that distinguishes between response patterns of poor and high performers; and second, to estimate a model for the latent trait. Furthermore, the results from the initial NRM can inform constraints to be used in a subsequent NRM. Then, the constrained model is used to estimate latent trait scores (θ) for each observation via discrimination and difficulty parameters (Bock, 1972; Thissen et al., 2010). In the original development of NRM, the resulting NRM-estimates from this process considerably improved score precision among respondents with 2PL scores below the median. In other words, using NRM IRT to estimate scores had advantages for poorer performing respondents (Bock, 1972).

Following suit, we use NRM IRT to both determine distractor order and estimate scores for our latent trait, logical reasoning using IRTPRO software (Cai et al., 2011). Prior to estimating the NRM, we assigned values to each answer choice based on text by Thissen et al. (2010). Within each item, the correct answer choice is assigned the highest value, and distractors are coded numerically for the remaining lower values in alphabetical order, with the minimum assigned distractor value has a parameterization value of 0. For example, Question 1 has six multiple choice options. We assign values of 0 through 4 to the five distractor answer choices in alphabetical order, so choice A as 0, choice B as 1, choice C as 2, choice E as 3, and choice F as 4. We then assign to the correct answer choice, D, the highest value, which is 5.

We estimate an initial NRM, allowing all discrimination and difficulty parameters to vary. This model has a marginal reliability of 0.48. Within each item, non-zero discrimination parameters are relative to the zero-parameterized choice. We decide on distractor ordering based on this first NRM (Thissen et al., 2010). For example, Table 3 shows the original discrimination parameters for Question 1 in Column 1. As expected, the correct choice D has the highest discrimination value relative to choice A. Choice C has the lowest discrimination value relative to choice A. Discrimination parameters for choices B, E, and F are similar in magnitude and much larger than choice A or C. Relative to A and C, B, E, and F also have visual properties that are incorrect to a similar degree. Therefore, we

constrain discriminating parameters for choices B, E, and F to the same value in a new model. Constrained NRM discrimination parameters are in column 2 of Table 3 which shows that choices B, E, and F all have the same values. Choice A is parameterized to 0, while choice C and choice D continue to have the lowest and highest values, respectively.

Table 3
Question 1 discrimination parameters for Malawi youth sample (N=2,513)

Response choices	Unconstrained NRM	NRM with constraints
A	0.00	0.00
B	0.77	0.73
C	-0.40	-0.44
D	1.60	1.50
E	0.87	0.73
F	0.80	0.73

Note: Using the first test item as an example, this table reports the estimated discrimination parameter values from both unconstrained and constrained NRM models. The column 1 model informed the constraints used in column 2 model, which were that choices B, E, and F would have equal discrimination parameter values.

Similar decision-making processes are conducted for each of the remaining seven items. For Questions 3, 6, and 7, the correct response does not have the highest discrimination parameter. Therefore, we group the correct response with distractors of similar or higher parameter values. This ensures latent scores for individuals choosing the correct response for those items are not 'penalized' relative to participants choosing a distractor with a higher discrimination parameter value. We allow the difficulty parameters—which essentially reflect the sample proportion endorsing a single response—for all response choices to vary within each grouping.

For easy comparison across the three methods estimating logical reasoning—summed scores, 2PL IRT, and constrained NRM IRT—we transform each score type to be normally standardized. Figure 1 consists of the three histograms showing the standardized distributions. Latent scores are estimated from the constrained model for each respondent. The resulting distribution in right panel of Figure 1 is direct evidence of the improvement, as there is very little skew (p=0.15).

The Zambia Study

Data and sample description

Data for the caregiver sample comes from the Government of Zambia's Child Grant Program impact evaluation also conducted by The Transfer

Project. The program targeted all households with a child under three years old living in the rural areas of the Kaputa, Shangombo and Kalabo districts. The evaluation was a cluster-randomized controlled trial with treatment and control groups. Randomized treatment assignment was established at the community level (Seidenfeld & Handa, 2011). The same eight previously described Raven's items and test administration protocol used in Malawi were included in Zambia during the program's 30-month follow-up survey (The Transfer Project, 2013b). Because these items were not included in data collection periods prior to program exposure, we exclude caregivers living in communities assigned to the treatment group to eliminate any potential differences in performance attributed to the program. Of the 2,360 caregivers participating in the 30-month follow-up survey in 2013, 1,173 caregivers living in control communities are included in this analysis. All caregivers are female, the mean age of the analysis sample is 32.74 years, and 29.24 percent have never been to school (Table 4). While households in the Zambia study also live in extreme poverty (average daily consumption is US$0.35 per person, similar to households in the Malawi sample), the demographic profile is very different, which allows us to assess whether our approach provides similar results among different subjects.

Table 4
Sample characteristics of caregivers in Zambia (N=1,173)

	Mean	Standard deviation
Age (years)	32.74	9.80
Educational attainment		
No primary	0.29	0.46
Some primary: grades 1-6	0.45	0.50
Completed primary: grade 7	0.15	0.35
Some secondary: grades 8-11	0.10	0.29
Completed secondary	0.01	0.12
Household characteristics		
Member(s) with disability	0.08	0.28
HH expenditures per capita, ZMW	53.38	58.35

Note: This table reports the mean value for each descriptive variable for caregivers in Zambia.

We follow the same procedure for estimating scores for Zambia conducted for Malawi: first examining summed score distributions, proceeding with 2PL IRT model fitting and score estimation, determining distractor ordering via an unconstrained NRM IRT, and finally, estimating logical reasoning scores using a constrained NRM. The overall purpose is to assess whether the IRT approach can successfully be used across very different study populations, which strengthens the external validity of the

method, but direct comparisons between Malawi and Zambia samples do not make sense because of their inherent differences.

Table 5
Question 1 discrimination parameters for the Zambia caregiver sample (N=1,173)

Response choices	Unconstrained NRM	NRM with constraints
A	0.00	0.00
B	0.56	0.57
C	-0.18	-0.20
D	0.57	0.57
E	0.15	0.16
F	0.00	0.00

Note: Using the first test item as an example, this table reports the estimated discrimination parameter values from both unconstrained and constrained NRM models in Zambia. The column 1 model informed the constraints used in column 2 model, which were that choices B and D would have equal discrimination parameter values and choices A and F would be parameterized to 0.

An example of discrimination parameters for item 1 from the constrained and unconstrained models is in Table 5. The NRM constraint configurations are unique to each country, as there might be inherent differences between youth in Malawi and caregivers in Zambia that could cause distinct response patterns within each sample.

Estimates of latent logical reasoning scores using IRT

Figure 3 shows the distributions for the standardized summed, 2PL, and constrained NRM scores for logical reasoning. Although there are three respondents who correctly answered all Raven's items, summed scores are extremely skewed (p<0.001), Table 6 shows over half of the respondents have summed scores of zero or one. Figure 4 also displays the average summed score for all multiple choice options for each Raven's item. The correct choice for each item has the highest average summed score, but among distractors, there is generally little variation. The summed score distribution is similar to what we observe in the Malawi sample—highly skewed with a very low average number of correct responses. The 2PL scores we estimate show improved variation; however, the largest proportion of participants still fall among the lowest range of values. As a result, the 2PL score distribution for Zambia is heavily skewed (p<0.001).

Figure 3. Comparing standardized distributions of logical reasoning among caregivers in Zambia.

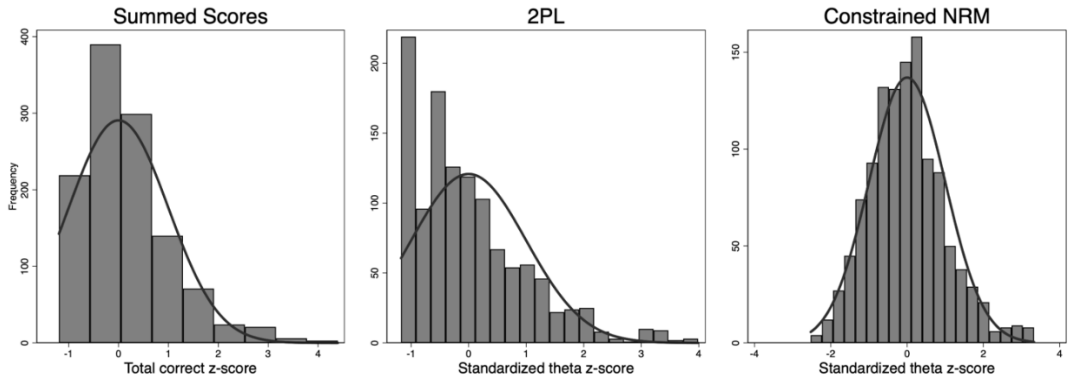


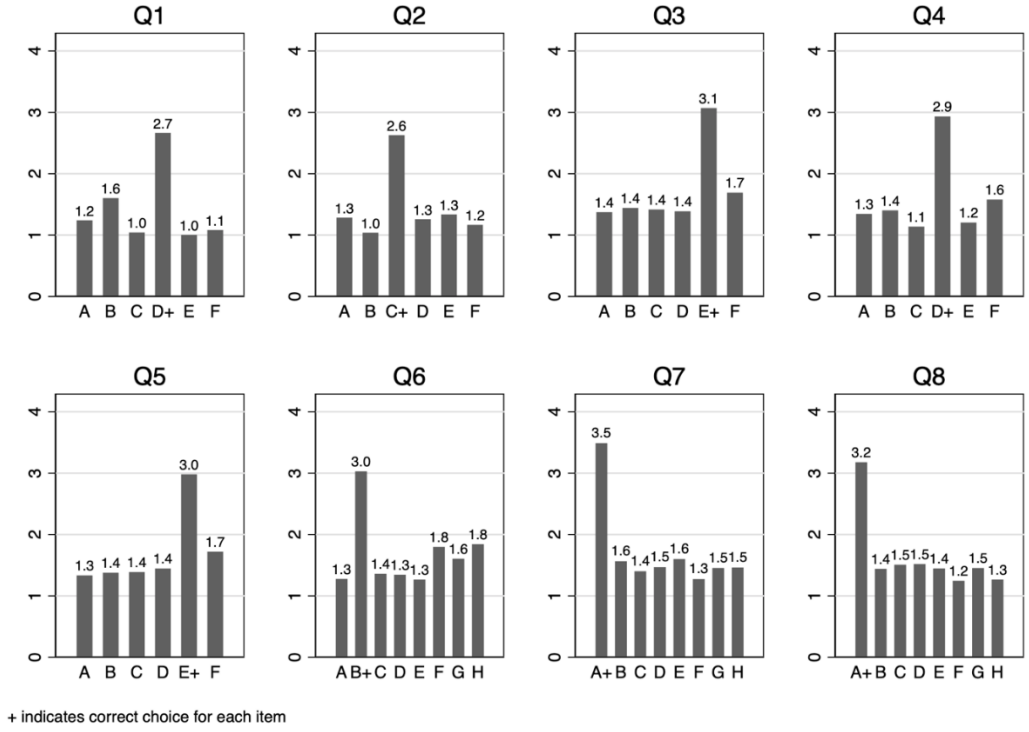
Table 6
Distribution of summed scores among caregivers in Zambia (N=1,173)

Summed score	N	%
0	221	18.72
1	392	33.22
2	201	35.51
3	141	11.95
4	71	6.02
5	24	2.03
6	21	1.78
7	6	0.51
8	3	0.25

Note: This table reports the number and percentage of Zambia caregivers across all possible summed score values, calculated by totaling the number of correct responses for each participant.

We then fit an initial NRM IRT to estimate the parameters to inform distractor ordering. This ordering determines the constraint structure for the subsequent NRM, which we use for score estimation and to maximize the variation among the lower end of the skew. Note that the three respondents who answered all eight items correctly have the same score because there is a single response pattern. Even though the tests for normality indicate the NRM estimated scores are still skewed (p<0.001), the constrained NRM logical reasoning score distribution in Figure 3 shows substantial improvements in the variation at the lower end. Thus, even in this adult sample, the proposed methodology can better distinguish nuance in the distribution of logical reasoning skills.

Figure 4. Mean sum score for each item response chosen among caregivers in Zambia.



+ indicates correct choice for each item

Correlates of the predicted logical reasoning scores

In multi-topic surveys like ones described earlier and the two used here, the purpose of including these short item Raven's tests is to obtain a measure of logical reasoning in analyses involving other behavioral characteristics. In this section we explain our procedure for exploring NRM-estimated logical reasoning score validity in two ways. First, we examine the determinants of the score, and then we assess the score's ability to predict behaviors of interest using Stata15 (*Stata/IC 15.1 for Mac*, 2018). As the purpose of this part of the analysis is to consider score validity, we also contextualize these results with existing literature. The discussion section more broadly considers the NRM IRT scoring approach for Raven's short sets and the score validation results across both samples. The analysis on the correlates of scores includes 2,487 youth with complete cases in Malawi (we exclude 26 youth due to missing information on the variables of interest) and 1,166 caregivers in Zambia (7 participants had missing data). Additional substantial reductions to the sample sizes are noted, some of which are due to questionnaire skip patterns.

First, we investigate a set of individual and household factors as predictors of logical reasoning. We examine three separate linear regression

models within each country. The three dependent logical reasoning variables are standardizations of the three score estimates (constrained NRM, 2PL, and summed) so that coefficients would be on similar scales. Note that, although the three measures are scaled similarly, comparing effect sizes between the two countries is inappropriate because each IRT-estimated score reflects only the population from which the score was estimated. Put simply, a youth participant from Malawi and a caregiver from Zambia with the same IRT-estimated score value cannot, and should not, be assumed to have the same logical reasoning skills, since scores are relative to the population from which they are estimated.

The individual characteristics we use as predictors in these models are gender as a binary variable and age as continuous, as well as educational attainment. The education variables are categorical to explore possible non-linear relationships, with no primary education is the referent group category, including those with only pre-primary or no education. Any individuals who had completed or attended higher than secondary education are collapsed into the highest education category. Other than the lowest and highest categories, education variables for Malawi and Zambia are constructed differently—based on basic characteristics of each sample—in order to maximize the ability to detect nuanced relationships when using logical reasoning scores as the model's dependent variable. The Malawi sample size is larger and has a higher proportion of participants who had attended school, so each grade level is a separate category. In Zambia, almost 30% of participants had no primary education, so it would have been very difficult to detect differences at the individual grade level. Therefore, we collapse categories based on the educational system: some primary for grades 1-6, completed primary for grade 7, some secondary for grade 8-11, completed secondary for grade 12 or higher.

The household-level covariates from both samples includes whether any members had disabilities and per capita expenditures (logarithmically-transformed). In Malawi, since many of the youth resided with their families and none were designated the main respondent (MR) to the broader household survey, we also used MR characteristics as covariates for those models, which include: MR literacy, gender, school attendance, and age. In Zambia, caregivers were the designated MR, so equivalent variables are perfectly collinear with the individual-level predictors previously mentioned.

We then estimate a second set of models using logical reasoning scores as predictors of a set of outcomes to check for predictive ability. We fit separate probit models predicted by the three versions of logical reasoning for a variety of dichotomous outcomes relevant to each sample. In Malawi, the dependent variables fall broadly under the topics of risk preferences and sexual experiences. For Zambia, we investigate logical reasoning as a predictor of different outcomes pertaining to future-oriented thinking.

From the probit models, we estimate the marginal change in the predicted probability of each outcome associated with a one unit change in each standardized logical reasoning score-type. Sample sizes, means, and standard deviations for all outcome variables are reported in Table 7.

The Malawi youth questionnaire assessed risk preferences through lottery preference items. Youth were given a hypothetical scenario in which they could choose to receive 3000 Malawian Kwacha (MWK) for certain, or flip a fair coin to either win 3000 MWK ('heads') or win nothing ('tails'). For respondents stating they would not flip the coin, the winning amount was increased to MWK 6000. Respondents who still would not choose to flip the coin were then asked to state the value of the winning amount that would

Table 7
Sample statistics for outcomes predicted by logical reasoning scores for Malawi and Zambia

	N	Mean	Standard deviation
Malawi			
Play lottery for 3000 MWK	2,487	0.60	0.49
Play lottery for 6000 MWK	998	0.58	0.49
Play lottery for >6000 MWK vs. ≤ 6000 MWK	2,487	0.17	0.38
Condom use at debut	793	0.36	0.48
Recent condom use	534	0.43	0.50
Zambia			
Always waits for future money	1,163	0.07	0.26
30 in a month ZMW	1,163	0.48	0.50
40 in a month ZMW	1,163	0.28	0.45
60 in a month ZMW	1,163	0.12	0.33
80 in a month ZMW	1,163	0.03	0.16
Often or always think about future when spending	1,173	0.50	0.50

Note: This table reports the sample sizes, means, and standard deviations for the variables we use as outcomes in models predicted by logical reasoning scores in Malawi and Zambia samples.

induce them to give up the guaranteed 3000 MWK and play the gamble. From these series of questions, we generate three dichotomized variables:

- Would play lottery for 3000 MWK (full sample)
- Would play lottery for 6000 MWK. This included only youth who would not play for 3000 MWK.
- Would only play for more than 6000 MWK (full sample)

The two sexual experiences variables includes only youth who had disclosed ever having sexual intercourse (N=797). The variables of interest

were whether youth reported having used a condom during their first intercourse and at their most recent encounter.

For Zambia, most outcomes are generated from a set of tasks measuring patience or intertemporal preferences. Respondents were asked about a series of scenarios under a hypothetical premise that they had suddenly won the lottery. In each scenario, participants had to choose between receiving the lottery winnings of 20 new Zambian Kwacha (ZMW) today or another amount in a month. The future amount takes the following differing values: 30, 40, 60 and then 80 ZMW. The future values are presented to respondents in a random order, so it is possible to observe 'double switches.' We exclude nine 'double-switch' cases as they may not have understood the questions correctly, choosing not to wait for a specific future value, and then chooses to wait for a lower future later in the sequence. The variables constructed from this intertemporal choice task are as follows:

- Present preference for all lottery questions (choosing 20 ZMW today for all four scenarios, never waits for future money)
- Switch points at each future value (30, 40, 60, 80), indicating the lowest value at which a respondent switched from present to future responses

Although we only present coefficients pertaining to logical reasoning scores as predictors of each model, we use some of the same controls as outlined in the models predicting logical reasoning, with a few notable changes. We construct education in both Malawi and Zambia using the categories previously described for Zambia. Education systems in Malawi and Zambia have the same grade structure. We combine grade categories for Malawi because the strong predictive relationship between education and logical reasoning—to be described in the next section—could mask effects of logical reasoning as a predictor of these other outcomes. Considering the substantial reduction to the number of observations for variables pertaining to sexual experiences, having separate categories for each grade could substantially diminish degrees of freedom. For similar reasons, we collapse secondary completion with some secondary, since few respondents had completed secondary education in either country.

Results

Predictors of Raven's scores

Estimates from models using the three Raven's scores as dependent variables in each country can be found in Table 8. Age is not significantly related to any type of logical reasoning score among Malawi or Zambia samples. In full Raven's tests, scores typically plateau in adolescence, regardless of score level (J. Raven, 2000). Given that neither sample

included participants younger than 12 years, our results align with what one expects given these populations' age ranges.

In Malawi, females have significantly lower scores in all three models than males. Gender differences have been found in other literature using RPMs to measure cognitive and reasoning skills. A meta-analysis including studies from a variety of countries found males had significantly higher RPM scores around age 14. The gap between male and female scores continued to grow throughout adolescence and into early adulthood (Lynn & Irwing, 2004).

There are positive relationships between education and logical reasoning for all score types in both Malawi and Zambia. Other studies of populations from LMICs demonstrate evidence that educational attainment is associated with better RPM performance (Jukes et al., 2018; Stein et al., 2005; Teivaanmäki et al., 2017). However, the direction of causality is not readily apparent. It is difficult to tell if more schooling has a positive impact on the score or if higher levels of logical reasoning facilitate better school performance thereby advancing educational attainment (Glewwe, 1991; Jamison & Lockheed, 1985).

In Malawi, the grade level at which such differences in scores are detected is different for each score type. Relative to respondents with no primary education, consistent significant differences in NRM-estimated logical reasoning scores begin in grade 4, with initial significant differences also occurring at grade 2. Differences in 2PL estimated-scores are steadily significant at grade 8 and above, with some evidence of score disparities at grade 6 and 4. Grade 6 is the point at which significant differences in summed scores become detectable. Thus, in Malawi, our analysis indicates that NRM-estimated scores detect differences in logical reasoning up to four grades lower than summed scores and two grades lower than 2PL-estimated scores.

Research by Van de Vijver and Brouwers' (2009), also conducted in Malawi, used summed scores of children ages seven to 14 responding to a select set of RPMs. The authors concluded that schooling does not enhance reasoning skills. Considering most children in their sample were not old enough to complete grade 6 (Van de Vijver & Brouwers, 2009), these findings are consistent with our summed score model results for Malawi, which also fails to distinguish differences in logical reasoning prior to grade 6. However, there are drastic age-related gains to skills in this same age range for RPM score distributions standardized for other populations (J. Raven, 2000). Even with NRM-estimated scores, it might be difficult to differentiate between the effects of natural development and the effects of education. However, in our sample of Malawi youth, the effect of education is likely more apparent because of the relative stability in overall cognition between ages 13 and 19 (J. Raven, 2000). For this reason, it is impossible to

Table 8
Multivariate effects on three versions of cognitive scores

		Malawi (N=2,487)			Zambia (N=1,166)		
		NRM	2PL	Summed score	NRM	2PL	Summed score
Female		-0.21**	-0.26**	-0.25**			
		(8.14)	(7.63)	(9.22)			
Age		-0.01	-0.01	-0.01	-0.00	-0.00	-0.00
		(0.54)	(0.61)	(0.77)	(0.40)	(1.18)	(0.93)
Education or grade level attained							
No primary/ none (referent)	0	--	--	--	--	--	--
Some primary	1	0.04	0.04	0.06	0.07	0.04	0.07
		(0.37)	(1.17)	(1.22)	(0.86)	(0.57)	(0.85)
	2	0.10*	0.01	0.03			
		(3.38)	(0.20)	(0.41)			
	3	0.10	0.02	0.05			
		(2.25)	(0.39)	(0.88)			
	4	0.29*	0.25*	0.26			
		(5.12)	(3.64)	(2.83)			
	5	0.41**	0.29	0.26			
		(10.43)	(3.04)	(2.67)			
	6	0.59***	0.41**	0.40**			
		(41.58)	(9.43)	(7.34)			
Completed primary	7	0.63**	0.40	0.37*	0.18	0.20*	0.23*
		(6.93)	(2.78)	(3.41)	(1.78)	(2.06)	(2.45)
Some secondary	8	0.96***	0.76**	0.77**	0.33**	0.32*	0.38**
		(26.97)	(6.33)	(9.05)	(2.72)	(2.60)	(2.94)
	9	0.91*	0.90*	0.87**			
		(5.83)	(5.39)	(6.91)			
	10	1.01**	0.96**	1.02**			
		(5.89)	(7.97)	(8.61)			
	11	1.40*	1.46*	1.54**			
		(4.10)	(4.76)	(5.87)			
Completed secondary	12+	1.34**	1.60**	1.43***	0.76*	0.64	0.75
		(10.53)	(7.69)	(14.99)	(2.46)	(1.68)	(1.86)
Household Indicators							
Household member(s) with disability		0.05	0.03	0.05	0.09	-0.00	0.02
		(1.44)	(0.74)	(1.10)	(0.79)	(0.04)	(0.17)
HH expenditures per capita [a]		-0.05	-0.03	-0.08	-0.17**	0.11	0.12*
		(1.09)	(0.64)	(2.19)	(3.18)	(1.96)	(2.42)
Female MR		0.15**	-0.03	-0.02			
		(8.75)	(2.12)	(1.46)			

(Table 8 continued)

	Malawi (N=2,487)			Zambia (N=1,166)		
	NRM	2PL	Summed score	NRM	2PL	Summed score
Literate (Chichewa or English) MR	0.02	0.02	0.07			
	(0.35)	(1.13)	(1.30)			
MR ever went to school	-0.02	-0.06	-0.03			
	(0.26)	(0.93)	(0.48)			
MR age	-0.00	-0.00	-0.00			
	(0.54)	(0.61)	(0.77)			
Y intercept	0.30	0.32	0.87	-0.70**	-0.37	-0.47*
	(0.80)	(0.62)	(2.54)	(3.01)	(1.64)	(2.18)
R^2	0.10	0.09	0.09	0.03	0.02	0.03

Note: *$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$; This table reports results from three separate multivariate linear regression models within each country predicting each type of logical reasoning score estimated from the RPMs. Columns represent distinct models. All score-related variables are standardized. For each row of covariates, T-statistics are inside parentheses under each associated coefficient estimate. [a] Models use logarithmic transformation of household expenditures per capita. MR refers to the survey main respondent.

directly compare the relationships we observe in the NRM-estimated model to Van de Vijver and Brouwers' results which use summed scores.

In Zambia, the age range is much wider, and thus, the educational experiences of participants are highly and immeasurably varied. Secondary school attendance (p=0.019) and completion (p=0.022) are significantly related to the NRM-estimated logical reasoning scores. Only some secondary education is significantly related to the 2PL-estimated scores, despite substantially higher coefficient values for the secondary completion category relative to coefficients for lower levels of education. NRM and 2PL-estimated score models detect no significant relationships between logical reasoning scores of respondents who completed primary and did not attend primary, but this relationship is significant in the summed score model. Because almost 15% of the sample completed primary school, we do not believe the lack of results we see for NRM and 2PL scores emanates from a power issue. Instead, the significant relationship we observe between primary completion and summed scores is likely due to 'overweighting' of certain items, since all items have equal contribution to the summed score (DeVellis, 2017; McNeish & Wolf, 2019), and limited variation from score construction resulting in underestimated standard errors. Summed scores are essentially ordinal data analyzed linearly, which has potential consequences in terms of misrepresenting the significance.

Most household-level factors are not related to any type of score. However, there are couple of exceptions. Having a female MR in Malawi is associated with significant increases to NRM-estimated scores relative to having a male MR. In Zambia, higher household expenditures is significantly associated with higher logical reasoning. Household

expenditures per capita have a stronger relationship with NRM estimated logical reasoning (0.17, p=0.004) than summed scores (0.12, p=0.021). This finding is consistent with other research showing associations between socioeconomic status (SES) and RPM scores (Hicks et al., 2017; Mani et al., 2013; J. Raven, 1989; Stein et al., 2005; Vogl, 2014). Other research in Zambia used 2PL to estimate scores from a subset of 10 Raven's items administered to heads of small-scale farming households and found a significant relationship between household asset quintile and Raven's score (Fehr et al., 2019). As with education, the direction of causality is not clear for the Zambia sample; logical reasoning skills may affect one's education, earning levels, and spending capacity.

Raven's scores as predictors

Exploring logical reasoning as a predictor of outcomes gives insight into how well the score types fit the context of existing research, providing evidence to support the validity of the additional variation captured by using both correct and incorrect responses in estimating logical reasoning through NRM IRT.

Table 9 presents coefficients for each logical reasoning score as a predictor of the outcomes explored for Malawi. In Malawi, two lottery outcomes were only significantly predicted by NRM-estimated logical reasoning, but not the summed score nor the 2PL. A standard deviation increase in NRM-estimated logical reasoning is associated with a 1.3 percentage-point higher probability of not tossing a coin for amounts less than 6000 MWK (p=0.003, column 3). Among youth who would not play the lottery for 3000 MWK, a standard deviation increase in NRM-estimated logical reasoning is associated with a 2.9 percentage-point lower probability of playing the lottery for 6000 MWK (p<0.001, column 2). There is no relationship between NRM scores and playing for 3000 MWK (column 1).

These lottery questions from Malawi measure risk preferences. The existing research on the relationship between cognition and risk preferences are mixed, but there is some work supporting our findings from this analysis. One study by Burks et al. using RPMs investigated risk-taking preferences and concluded subjects with higher Raven's scores better evaluated complex alternatives. Their choices were sensitive to the expected lottery values and therefore, participants with higher scores were neither necessarily risk-averse or risk-seeking (Burks et al., 2009). This aligns with what we observe in Malawi, where youth with higher logical reasoning scores are also more sensitive to the expected value of the lottery—they are less likely to risk playing the lottery to win 6000 MWK, but more likely to play for an amount over 6000 MWK. In contrast, the other measures are not sensitive to the expected value of the lottery. A meta-analysis by

Table 9
Logical reasoning as a predictor of outcomes among youth in Malawi

	Play lottery for 3000	Play lottery for 6000	Play lottery for >6000 vs. ≤ 6000	Condom use at debut	Recent condom use
Mean	0.60	0.58	0.17	0.36	0.43
NRM	-0.00	-0.03***	0.01**	0.02	0.05**
	(-0.34)	(-7.05)	(2.86)	(0.81)	(2.64)
2PL	0.00	-0.02	0.01	0.00	0.01
	(0.13)	(-0.94)	(1.16)	(0.14)	(0.64)
Summed	-0.01	-0.00	0.00	-0.00	0.02
score	(-0.31)	(-0.37)	(0.71)	(-0.00)	(0.91)
N	2,487	998	2,487	793	533

Note: * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$. This table reports the marginal probabilities from probit models of the variable indicated in the row on the outcome indicated at the top of each column, with z-statistics in parentheses below the coefficients. Models include additional control variables as described in the text. All score-related variables are standardized.

Lilleholt on risk aversion included studies using a variety of tools to measure "cognitive ability," including RPMs. The author concluded there is a weak relationship between ability and risk aversion in scenarios involving gains, but not in scenarios involving losses or those that are mixed (Lilleholt, 2019). The lack of relationship in loss scenarios supports our finding in Malawi that there is no significant relationship between logical reasoning and risk involving losses at 3000 MWK.

Table 10 shows the estimated relationships between logical reasoning as a predictor and the outcomes we investigate for the Zambia sample. A one standard deviation increase in NRM-estimated logical reasoning is associated with a 1.1 percentage-point increase in the probability of always choosing to wait for some future value (p=0.008, column 1), but 2PL and summed scores do not significantly predict this outcome. All three scores are significantly negatively related to switching from present to future preference at the highest increment of 80k ZMW, and all three are associated with often/always thinking about the future when spending money.

The lottery questions from Zambia are intended to measure preferences related to time. The aforementioned study by Burks et al., found higher levels of Raven's scores were associated with greater patience in an analysis using similar-type lottery scenarios as the ones from Zambia (Burks et al., 2009). Among a large sample of adults in Germany, scores from other assessments measuring cognition were negatively related to impatience, indicating individuals with higher scores were more likely to be patient

(Dohmen et al., 2010). Both of these pieces of research are consistent with our finding that respondents in Zambia with higher logical reasoning scores are significantly more likely to ever wait for future money, and thus are more patient.

Table 10
Logical reasoning as a predictor of outcomes among caregivers in Zambia

	Switch points					Often or always think about future when spending
	Always waits for future money	30 in a month	40 in a month	60 in a month	80 in a month	
Mean	0.02	0.48	0.28	0.12	0.03	0.50
NRM	0.01**	0.00	-0.00	-0.00	-0.01*	0.05**
	(2.67)	(0.07)	(-0.04)	(-0.08)	(-2.25)	(2.89)
2PL	0.01	-0.01	0.02	0.01	-0.01*	0.05**
	(1.42)	(-0.63)	(1.12)	(0.36)	(-2.04)	(2.89)
Summed	0.01	-0.00	0.01	0.01	-0.01*	0.05***
score	(1.65)	(-0.27)	(0.65)	(0.44)	(-2.19)	(3.53)
N	1,156	1,156	1,156	1,156	1,156	1,166

Note: * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$. Table reports the marginal probabilities from probit models of the variable indicated in the row on the outcome indicated at the top of each column, with z-statistics in parentheses below the coefficients. Models include additional control variables as described in the text. All score-related variables are standardized.

In both countries, we also include other variables capturing behaviors indirectly related to risk and intertemporal preferences. For example, condom use behavior may reflect an individual's predisposition to plan ahead and avoid large risk. In Malawi, there is no relationship between logical reasoning and condom use at first sexual encounter. We posit that the first, singular sexual encounter may not be indicative of one's preference to use a condom, as a variety of other factors could contribute to one's first sexual encounter, including levels of STI and pregnancy prevention knowledge and whether that encounter was chosen. We also look at condom use at the most recent sexual encounter, and find NRM-estimated logical reasoning is a significant predictor of this outcome. A one standard deviation increase in NRM-estimated logical reasoning is associated with a 5.2 percentage-point increase in the probability of recently using a condom (p=0.015), while the other two measures are not associated with recent condom use. Similarly, in the United States, adolescent girls categorized as

having low cognitive abilities—measured through a picture vocabulary test—were significantly more likely to not use birth control at either their first or most recent encounter relative to girls with "average" cognitive ability (Cheng & Udry, 2005). Other research has shown that summed scores from a short set of Raven's items were significantly related to openness to using a condom. When respondents were primed with information on STI prevention and given hypothetical scenarios on sexual encounters, summed scores were significantly associated with increased likelihood of condom use (Lee et al., 2020). Among a sample of adults from Peru, Raven's score was included as a variable in a structural equation model as one of the several measures of cognitive ability; cognitive ability was significantly and positively related to condom knowledge and use (Dieckmann et al., 2015). All of these findings are generally consistent with our results on logical reasoning's relationship with recent condom use among youth in Malawi. Given that knowledge on sexual health may be gained after and in reaction to one's initial encounter, these articles also support our theories as to why we only observe this relationship at the most recent encounter.

Discussion

In this population of Malawi youth and Zambian caregivers living in extremely poor households, the summed score distributions from the select Raven's items are extremely skewed. From both theoretical and practical perspectives, the skewed summed score distribution is problematic. It questions the theoretical appropriateness of treating summed Raven's scores as a continuous measure, and thus, its linearity. From a practical perspective, we cannot be sure there is enough variation to pursue regression models using continuous summed scores. NRM IRT-estimated scores offer an advantage by increasing the estimated variation through the use of information contained in incorrect responses, raising fewer questions on the appropriateness of treating the latent variable as continuous. Additionally, since individual scores are standard deviations of logical reasoning, it can be easily transformed into percentiles, which provides interpretative benefits as well.

With the exception of one education category in the Zambia sample, there are no predictor or outcome variables significantly associated with summed or 2PL scores, but not significantly associated with NRM scores. On the contrary, there are several significant relationships when using NRM scores that are not significant for 2PL and summed scores. Relative to 2PL or summed scores, NRM estimation was better able to predict variation in risk preferences, condom use, and patience. In terms of construct validity, lower and higher levels of education were better able to predict variation in the NRM score relative to 2PL and summed scores. These findings suggest

that, when relationships are less strong, NRM can uncover differences that are not apparent when using 2PL or summed scores. Thus, these relationships are a reflection of the improved nuance and smoother distribution we gain by estimating logical reasoning using NRM IRT.

As we can see from this analysis, scoring method choice has consequences in terms of conclusions drawn from subsequent analyses (McNeish & Wolf, 2019). This analysis highlights the potential substantial implications stemming from scoring estimation choice. Comparing across the three models, the NRM-estimated scores strike a balance, reflecting patterns qualitatively similar to the sum score model while also allowing us to take advantage of the previously described distributional and interpretative benefits of IRT more generally. For the variables where the NRM-estimated scores show improved sensitivity as a predictor and an outcome, our findings align with other literature investigating similar relationships, supporting our claim that the additional nuance recovered is valid.

Importantly, this analysis demonstrates a way to standardize RPMs to the populations in which the tests are conducted. Many researchers acknowledge the limitations of applying British norms to score RPM tests in other non-British or non-Western contexts. While the origins of the RPMs will always be inescapably British, using NRM to inform choice ordering upon which scores are estimated reduces the cultural bias from which the test was created without fundamentally changing the test itself. In other words, maximizing the inductive process maximizes both analytical impartiality and a study population's representation in the estimation procedure. Thus, cultural bias embedded in ensuing conclusions drawn from analyses using scores estimated by this procedure should theoretically be minimized.

This NRM-based procedure has applications beyond RPMs and could be used to estimate scores from other multiple choice assessments. However, the improvements to score nuance and variation—generated through the NRM process—have limits, especially if tests are too easy. The distribution of summed scores in our analysis are low, indicating the test was very difficult for both samples. The additional nuance we exploit through NRM IRT comes from incorrect responses. In contrast, a test that is too easy would have a large proportion of summed scores at the highest end of the distribution. By definition, all participants with perfect summed scores have the same response pattern by only choosing correct responses (assuming each test question only has one right answer). Tests that are too easy, consequently, have large proportions of respondents with perfect performance. Therefore, there is no variation in the response patterns to exploit among these participants. The relative sample-wide advantage of the NRM procedure over 2PL might be minimal in terms of uncovering variation in wrong responses when tests are too easy. This notion is

supported by Bock's initial paper describing NRM, showing that precision was similar among respondents with 2PL scores above the median (Bock, 1972).

There are other optimal conditions for using information in both right and wrong answers to estimate latent scores from tests. First, scoring using wrong responses only helps to the extent that there actually are wrong responses (Bock, 1972; Thissen, 1976). As is the case for most statistical strategies, this process also works best for larger samples sizes compared with smaller ones. Additionally, the NRM ordering and estimation procedure is probably most beneficial for short tests. Tests with a large number of items are more reliable. For example, if the test is not too difficult, a set of RPMs containing 36 items with responses categorized as dichotomous right-wrong has more opportunities to capture variation compared to an eight item test. In this case, the 2PL scores may be sufficient and NRM estimation may be excessive. However, in multi-topic population-based surveys such as the ones we examine here, there is typically not enough room for a long forms of RPM tests.

At the risk of sounding cliché, our strengths in this analysis are also our limitations since we pursued this procedure to overcome basic challenges in the data. The limited set of RPM items, in conjunction with their relatively high difficulty, resulted in skewed distributions. And while the multivariate models and the improved marginal reliability with the NRM estimated scores provide some degree of confidence there is a latent construct we presume to be logical reasoning, there is an inherent risk of overfitting the data with any exploratory, inductive process. Our described procedure is not immune to that risk.

Another limitation of this analysis pertains to the ordering of the distractors. Although ordering was largely informed by the initial NRM results, some element of subjective reasoning was necessary for deciding which items would be constrained together. It would have been ideal to have members of the study population review or participate in these decisions to further remove researchers' cultural bias from the analytical procedure.

Conclusion

After applying a strategy using two iterations of NRM IRT in each of two distinct samples, the distribution of latent logical reasoning scores are much improved. Because NRM IRT packages are now commonly available in statistical software such as Stata (StataCorp, 2023), SAS (SAS Institute Inc, 2017), and R (Chalmers, 2023), this estimation strategy is more widely accessible to a spectrum of disciplines. We propose this new method be used as an alternative to sum score strategies for estimating ability from short-item RPMs imbedded into multi-topic surveys. Doing so would likely result

in statistically sounder models that are more culturally representative of populations being studied.

Author notes: Inquiries can be directed to the first author, Hannah Silverstein, at 135 Dauer Drive 401 Rosenau Hall, CB #7445 Chapel Hill, NC 27599-7445; phone: 336-686-4636, email: Silver24@live.unc.edu. Author contributions: HS conceived of the presented idea, performed the analyses, drafted the manuscript, and designed the figures. SH is the principal investigator of the Transfer Project from which data originated. DT helped implement and verify the analytical methods. SH and DT aided in interpreting the results, worked on the manuscript, and supervised the findings of this work. All authors discussed the results and contributed to the final manuscript.

References

Abdoulayi, S., Angeles, G., Barrington, C., Brugh, K., Handa, S., Hill, M. J., Kilburn, K., Otchere, F., Zuskov Csr-Unima:, D., Mvula, P., Tsoka, M., & Unima, C. (2014). *Malawi Social Cash Transfer Program Baseline Evaluation Report*. https://transfer.cpc.unc.edu/wp-content/uploads/2021/04/Malawi-SCTP-Baseline-Report.pdf

Akresh, R., De, D., Harounan, W., The, K., & Bank, W. (2013). *Cash Transfers and Child Schooling Evidence from a Randomized Evaluation of the Role of Conditionality Impact Evaluation Series No. 82*. https://documents.worldbank.org/en/publication/documents-reports/documentdetail/587731468005971189/cash-transfers-and-child-schooling-evidence-from-a-randomized-evaluation-of-the-role-of-conditionality

Alkire, S., & Samman, E. (2014). Mobilising the Household Data Required to Progress toward the SDGs. *OPHI Working Paper 72*, *September*, 1–51. http://www.ophi.org.uk/mobilising-the-household-data-required-to-progress-toward-the-sdgs/

Azubuike, O. B., Briones, K., & Lives, Y. (2016). *Young Lives Rounds 1 to 4 Constructed files*. www.younglives.org.uk

Beaman, L., & Magruder, J. (2012). Who gets the job referral? Evidence from a social networks experiment. *American Economic Review*, *102*(7), 3574–3593. https://doi.org/10.1257/aer.102.7.3574

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29–51. https://doi.org/10.1007/BF02291411

Burks, S. V, Carpenter, J. P., Goette, L., & Rustichini, A. (2009). Cognitive skills affect economic preferences, strategic behavior, and job attachment. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(19), 7745–7750. https://doi.org/10.1073/pnas.0812360106

Cai, L., Thissen, D., & Toit, S. du. (2011). *IRTPRO for Windows* (2.1). Scientific Software International.

Chalmers, P. (2023). *Package 'mirt.'* https://cran.r-project.org/web/packages/mirt/mirt.pdf

Charness, G., Rustichini, A., & van de Ven, J. (2018). Self-confidence and strategic behavior. *Experimental Economics*, *21*(1), 72–98. https://doi.org/10.1007/s10683-017-9526-3

Cheng, M. M., & Udry, J. R. (2005). Sexual Experiences of Adolescents with Low Cognitive Abilities in the U.S. *Journal of Developmental and Physical Disabilities*, *17*(2). https://doi.org/10.1007/s10882-005-3686-3

De Groot, R., Handa, S., Park, M., Darko, R. O., Osei-Akoto, I., Bhalla, G., & Ragno, L. P. (2015). *Heterogeneous Impacts of an Unconditional CashTransfer Programme on Schooling: Evidence from the Ghana LEAP Programme*. https://www.unicef-irc.org/publications/pdf/iwp_2015_10.pdf

DeVellis, R. F. (2017). Chapter 2: Understanding the latent variable. In L. Bickman & D. J. Rog (Eds.), *Scale Development: Theory and Applications* (4th ed.). Sage.

Dieckmann, N., Peters, E., Leon, J., Benavides, M., Baker, D., & Norris, A. (2015). The Role of Objective Numeracy and Fluid Intelligence in Sex-Related Protective Behaviors. *Current HIV Research*, *13*(5), 337–346. https://doi.org/10.2174/1570162x13666150511123841

Dohmen, T., Falk, A., Huffman, D., & Sunde, U. (2010). Are risk aversion and impatience related to cognitive ability? *American Economic Review*, *100*(3), 1238–1260. https://doi.org/10.1257/aer.100.3.1238

Dramé, C., & Ferguson, C. J. (2019). Measurements of Intelligence in sub-Saharan Africa: Perspectives Gathered from Research in Mali. *Current Psychology*, *38*(1), 110–115. https://doi.org/10.1007/s12144-017-9591-y

Dupas, P. (2011). Do Teenagers Respond to HIV Risk Information? Evidence from a Field Experiment in Kenya. *American Economic Journal: Applied Economics*, *3*(3), 1–34. http://www.jstor.org/stable/25760244

Dupas, P., & Robinson, J. (2013). Savings constraints and microenterprise development: Evidence from a field experiment in kenya. *American Economic Journal: Applied Economics*, *5*(1), 163–192. https://doi.org/10.1257/app.5.1.163

Fehr, D., Fink, G., & Jack, K. (2019). Poverty, Seasonal Scarcity, and Exchange Asymmetries. In *NBER Working Paper Series* (No. 26357). https://doi.org/10.1017/CBO9781107415324.004

Glewwe, P. (1991). *Schooling, Skills, and the Returns to Government Investment in Education: An exploration using Data from Ghana* (No. 76; Living Standards Measurement Study). https://files.eric.ed.gov/fulltext/ED335389.pdf

Hanaki, N., Jacquemet, N., Luchini, S., & Zylbersztejn, A. (2016). Fluid intelligence and cognitive reflection in a strategic environment: Evidence from dominance-solvable games. *Frontiers in Psychology*, *7*(AUG), 1188. https://doi.org/10.3389/fpsyg.2016.01188

Hicks, J. H., Kleemans, M., Li, N. Y., Miguel, E., Albouy, D., Alvarez, J., Beaman, L., Donovan, K., Faber, B., Finan, F., Gaubert, C., Gollin, D., Jayachandran, S., Johnson, T., Kaur, S., Lagakos, D., Morten, M., Mobarak, M., Mueller, V., … Hall, E. (2017). *Reevaluating agricultural productivity gaps with*

longitudinal microdata (No. 23253; NBER Working Paper Series). http://www.nber.org/data-appendix/w23253

Jamison, D. T., & Lockheed, M. E. (1985). *Participation in Schooling: Deferminants and Learning Outcomes In Nepal Education and Training Departrnent Operations Policy Staff* (No. EDT9; Education and Training Series). http://documents.worldbank.org/curated/en/393231468775752894/pdf/multi-page.pdf

Jukes, M. C. H., Zuilkowski, S. S., & Grigorenko, E. L. (2018). Do Schooling and Urban Residence Develop Cognitive Skills at the Expense of Social Responsibility? A Study of Adolescents in the Gambia, West Africa. *Journal of Cross-Cultural Psychology*, *49*(1), 82–98. https://doi.org/10.1177/0022022117741989

Lee, S. T. H., Li, N. P., Meltzer, A. L., Melia, N. V., & Oh, H. S. (2020). No glove, no love: General intelligence predicts increased likelihood of condom use in response to HIV threat. *Personality and Individual Differences*, *157*. https://doi.org/10.1016/j.paid.2020.109813

Lilleholt, L. (2019). Cognitive ability and risk aversion: A systematic review and meta analysis. *Judgment and Decision Making*, *14*(3), 234–279. https://doi.org/10.1017/S1930297500004307

Lindwall, M., Berg, A. I., Bjälkebring, P., Buratti, S., Hansson, I., Hassing, L., Henning, G., Kivi, M., König, S., Thorvaldsson, V., & Johansson, B. (2017). Psychological health in the retirement transition: Rationale and first findings in the HEalth, Ageing and Retirement Transitions in Sweden (HEARTS) study. *Frontiers in Psychology*, *8*(September). https://doi.org/10.3389/fpsyg.2017.01634

Lynn, R., & Irwing, P. (2004). Sex differences on the progressive matrices: A meta-analysis. *Intelligence*, *32*(5), 481–498. https://doi.org/10.1016/j.intell.2004.06.008

Mani, A., Mullainathan, S., Shafir, E., & Zhao, J. (2013). Poverty impedes cognitive function. *Science*, *341*(6149), 976–980. https://doi.org/10.1126/science.1238041

Martin, G. H., & Pimhidzai, O. (2013). *Service Delivery Indicators: Kenya, July 2013*. https://www.sdindicators.org/sites/sdi/files/SDI-Report-Kenya .pdf

McNeish, D., & Wolf, M. G. (2019). *Thinking twice about sum scores*. *41*, 1–50. https://doi.org/https://doi.org/10.31234/osf.io/3wy47

Molina, E., & Martin, G. (2015). *Education Service Delivery in Mozambique, October 2015*. https://catalog.ihsn.org/index.php/catalog/6918/download/82211

Pimhidzai, O., & Martin, G. (2015). *Education Service Delivery in Nigeria, October 2015*. http://microdata.worldbank.org/index.php/catalog/2752/download/39283

Raven, J. (1989). The Raven Progressive Matrices: A Review of National Norming Studies and Ethnic and Socioeconomic Variation Within the United States. *Journal of Educational Measurement*, *26*(1), 1–16. https://doi.org/10.1111/j.1745-3984.1989.tb00314.x

Raven, J. (2000). The Raven's Progressive Matrices: Change and Stability over Culture and Time. *Cognitive Psychology*, *41*, 1–48.

https://doi.org/10.1006/cogp.1999.0735
Raven, J. (2008). General Introduction and Overview: The Raven Progressive Matrices Tests: Their Theoretical Basis and Measurement Model. In J. Raven & J. Raven (Eds.), *Uses and Abuses of Intelligence: Studies Advancing Spearman and Raven's Quest for Non-arbitrary Metrics* (pp. 17–68). Royal Fireworks Press. https://www.researchgate.net/publication/255605513_The_Raven_Progressive_Matrices_Tests_Their_Theoretical_Basis_and_Measurement_Model#fullTextFileContent
Raven, J. C., Court, J. H., & Raven, J. (1986). Section 2: Coloured Progressive Matrices. In *Manual for Raven's Progressive Matrices adn Vocabulary Scales* (1968th ed.). Lewis.
Raven, J. C., Court, J. H., & Raven, J. (1992). Section 3: Standard progressive matrices. In *Manual for Raven's progressive matrices and vocabulary scales*. Oxford Psychologists Press.
Rubalcava, L., & Teruel, G. (2006). *Mexican Family Life Survey, First Round* (Working Paper). www.ennvih-mxfls.org
Rubalcava, L., & Teruel, G. (2008). *Mexican Family Life Survey, Second Round* (Working Paper). www.ennvih-mxfls.org
Rubalcava, L., & Teruel, G. (2013). *Mexican Family Life Survey, Third Round* (Working Paper). www.ennvih-mxfls.org.
SAS Institute Inc. (2017). *SAS/STAT 14.3 ® User's Guide The IRT Procedure*. https://support.sas.com/documentation/onlinedoc/stat/143/irt.pdf
Seidenfeld, D., & Handa, S. (2011). *Zambia's Child Grant Program: Baseline Report Final*. https://www.air.org/sites/default/files/downloads/report/Zambia_Child_Grant_Baseline_Report_2011.pdf
Stata/IC 15.1 for Mac (Revision 15 Oct 2018). (2018). StataCorp LLC.
StataCorp. (2023). STATA Item Response Theory Reference Manual, Release 18. In *Stata Reference Manual*. https://www.stata.com/manuals/irt.pdf
Stein, A. D., Behrman, J. R., DiGirolamo, A., Grajeda, R., Martorell, R., Quisumbing, A., & Ramakrishnan, U. (2005). Schooling, educational achievement, and cognitive functioning among young Guatemalan adults. *Food and Nutrition Bulletin*, *26*(2 SUPPL. 1), S46–S54. https://doi.org/10.1177/15648265050262s105
Steinberg, L., & Thissen, D. (2013). Item Response Theory. In J. Comer & P. Kendall (Eds.), *The Oxford handbook of research strategies for clinical psychology* (pp. 336–373). Oxford University Press.
Strauss, J., Beegle, K., Sikoki, B., Dwiyanto, A., Herawati, Y., & Witoelar, F. (2009). The Fourth Wave of the Indonesia Family Life Survey: Overview and Field Report. *RAND Working Paper*, *WR-675/1-N*(April 2009), i–82.
Strauss, J., Beegle, K., Sikoki, B., Dwiyanto, A., Herwati, Y., Witoelar, F., & Corporation, R. (2004). *The Third Wave of the Indonesia Family Life Survey (IFLS3): Overview and Field Report*. *WR-144/1-NIA/NCHID*.
Strauss, J., Witoelar, F., & Sikoki, B. (2016). *The Fifth Wave of the Indonesia Family Life Survey: Overview and Field Report: Volume 1*. www.rand.org/giving/contribute
Teivaanmäki, T., Bun Cheung, Y., Pulakka, A., Virkkala, J., Maleta, K., & Ashorn,

P. (2017). Height gain after two-years-of-age is associated with better cognitive capacity, measured with Raven's coloured matrices at 15-years-of-age in Malawi. *Maternal and Child Nutrition*, *13*(2), 1–12. https://doi.org/10.1111/mcn.12326

The Transfer Project. (2013a). *Young Person's Questionnaire, Baseline*. 1–4. https://transfer.cpc.unc.edu/wp-content/uploads/2015/10/Malawi_SCT_Young-Person-Questionnaire_Baseline.pdf

The Transfer Project. (2013b). *Zambia Social Protection Scheme Child Grant 30 Month Follow-up Survey 2013 Kalabo, Kaputa and Shang'ombo Districts*. https://transfer.cpc.unc.edu/wp-content/uploads/2015/10/Zambia_CGP_Household-Questionnaire_Follow-up_30-mo.pdf

Thissen, D. (1976). Information in wrong responses to the Raven Progessive Matrices. *Journal of Educational Measurement*, *13*(3), 201–214. https://doi.org/10.1111/j.1745-3984.1976.tb00011.x

Thissen, D., Cai, L., & Bock, R. D. (2010). The Nominal Categories Item Response Model. In *Handbook of Polytomous Item Response Theory Models*. https://doi.org/10.4324/9780203861264.ch3

Thissen, D., & Steinberg, L. (2009). Item Response Theory. In R. Millsap & A. Maydeu-Olivares (Eds.), *The SAGE Handbook of Quantitative Methods in Psychology* (pp. 148–177). SAGE Publications Ltd. https://doi.org/10.4135/9780857020994.n7

Tuan Pham Thi Lan, T. (2003). *Young Lives Preliminary Country Report: Vietnam*. www.younglives.org.uk

UNICEF. (2016). *Cognitive Capital: Investing in children to generate sustainable growth*.

Van de Vijver, F. J. R., & Brouwers, S. A. (2009). Schooling and basic aspects of intelligence: A natural quasi-experiment in Malawi. *Journal of Applied Developmental Psychology*, *30*(2), 67–74. https://doi.org/10.1016/j.appdev.2008.10.010

Vogl, T. S. (2014). Height, skills, and labor market outcomes in Mexico. *Journal of Development Economics*, *107*, 84–96. https://doi.org/10.1016/j.jdeveco.2013.11.007

Wane, W., & Martin, G. (2016). *Education Service Delivery in Uganda, March 2016*. http://pubdocs.worldbank.org/pubdocs/publicdoc/2016/3/812821457978473769/Session-8-Deon-Filmer.pdf

Wane, W., & Rakotoarivony, R. A. (2017). *Education Service Delivery in Madagascar*. https://www.sdindicators.org/sites/sdi/files/8. 2016 Madagascar_SDIEducation_FinalTechnicalReport.pdf