Norwegian University of Life Sciences (NMBU)

Numeracy Skills, Decision Errors, and Risk Preference Estimation

Stein T. Holden and Mesfin Tilahun

Norwegian University of Life Sciences
Centre for Land Tenure Studies

Centre for Land Tenure Studies Working Paper 05/23
ISBN: 978-82-7490-316-6

Numeracy Skills, Decision Errors, and Risk Preference Estimation

Stein T. Holden[1*] and Mesfin Tilahun[1,2]

[1*]School of Economics and Business, Norwegian University of Life Sciences, P. O. Box 5003, 1432 Ås, Norway.
[2]Department of Economics, Mekelle University, Mekelle, P. O. Box 451, Tigray, Ethiopia.

*Corresponding author(s). E-mail(s): stein.holden@nmbu.no;
Contributing authors: mesfintilahungelaye@gmail.com; mesfin.telahun.gelaye@nmbu.no;

Abstract

Basic numeracy skills are obviously important for rational decision-making when agents are facing choices between risky prospects. Poor and vulnerable people with limited education and numeracy skills live in risky environments and have to make rational decisions in order to survive. How capable are they to understand and respond rationally to economists' tools for the elicitation of risk preferences? Can we make designs that are simple enough for them to give rational responses that reveal their true preferences? And how much does variation in their limited numeracy skills contribute to decision errors and the estimated sizes of their risk preference parameters? Finally, we ask whether Expected Utility (EU) theory is sufficient or whether Rank Dependent Utility (RDU) does better in the analysis of decision errors and risk preferences in our context. We try to answer these research questions based on a large sample of rural youth business group members from Ethiopia based on two variants of a Certainty Equivalent - Multiple Choice List (CE-MCL) approach with 12 and 10 Choice Lists (CLs) per subject. Numeracy skill scores are constructed based on a math test with 15 contextualized questions. The experiment facilitates the estimation of structural models while separating the effects of numeracy skills on decision errors in a Fechner error specification that is a function of numeracy skills and experimental design characteristics. The structural models estimate alternatively Expected Utility (EU) and Rank Dependent Utility (RDU) models, the latter with two-parameter Prelec probability weighting functions. It allows us to assess whether limited numeracy skills are correlated with EU-type risk tolerance (utility curvature) and RDU-type of probabilistic risk tolerance in the form of probabilistic insensitivity and optimism/pessimism bias. We find that weak numeracy skills are associated with slightly less risk tolerance in EU models, with stronger probabilistic insensitivity in RDU models, and with more random noise (Fechner error) in both types of models. However, even the subjects with the weakest numeracy skills performed quite well in the simple CE-MCL experiments with the binary choice elicitation approach, indicating that it was capable of revealing the risk preferences of such subjects with very low numeracy skills as they produced only marginally more decision errors than subjects with better numeracy skills.

Keywords: Numeracy skills, Risk preferences, Field experiment, Ethiopia

JEL Classification: C93 , D81

1 Introduction

Numeracy skills are essential for human behavior, survival, and welfare in a world characterized by many types of risks and uncertainty. Intuitive numeracy skills are learnt through repeated interactions with real world situations through cognitive processes and interpretations of experiences (adaptive heuristic strategies). Basic and advanced numeracy skills are more efficiently learnt through schooling but there is large variation in such skills in developing as well as developed countries due to differences in cognitive ability, motivation, quality of education, job situation, and self-training (Garcia-Retamero, Sobkow, Petrova, Garrido, & Traczyk, 2019; Zhang & Holden, 2023). Financial literacy is essential for people doing business and requires statistical numeracy skills (Cokely et al., 2018). We study the numeracy skills and risk preferences of members of youth business groups with limited basic education that attempt to establish sustainable rural livelihoods in a semi-arid risky environment in Ethiopia.

There exists mixed evidence regarding the relationship between cognitive ability and risk preferences (Dohmen, Falk, Huffman, & Sunde, 2018; Lilleholt, 2019). The mechanisms underlying the relationship between cognitive abilities and individual risk preferences remain unclear (Mechera-Ostrovsky, Heinke, Andraszewicz, & Rieskamp, 2022). One possible mechanism is that those with higher cognitive ability have more confidence in judging risky options and making better decisions and are, therefore, more willing to take risks. On the other hand, people with lower cognitive ability may depend on cruder heuristics in their decision-making and this may correlate spuriously and unpredictably with the true risks and their underlying true risk preferences. When researchers use different elicitation devices to uncover the underlying true risk preferences, the cognitive abilities of the subjects also influence how successful the researchers are in achieving this. Can devices that facilitate the observation of subject-level decision errors, and how these vary with the elicitation device design characteristics and cognitive skills, have the potential to provide less biased estimates of risk preferences? Or do we need to resort to very simple designs that cannot separate decision errors from the true risk preferences when we study the risk preferences of subjects with limited education and numeracy skills? We test an experimental design and elicitation approach that is sophisticated enough to separate decision errors from preferences, and whether it is simple enough for even subjects with very limited numeracy skills to be able to give predictable responses.

To our knowledge there exist no comprehensive studies of numeracy skills and how they are related to decision errors and risk preferences of subjects with limited education in a developing country context. A study that relates numeracy skills to decision errors and risk preferences and includes subjects with limited numeracy skills in Canada is Dave, Eckel, Johnson, and Rojas (2010). They compare two different risk elicitation methods, the more complex and cognitively more demanding Holt and Laury (2002) (HL) Choice List approach and the simpler Eckel and Grossman (2002, 2008) (EG) approach[1] combined with a survey to measure mathematical skills. The much-used HL approach has a Choice List (CL) that contains 10 choices between two risky prospects with probabilities ranging from 0.1 to 0.9. The EG approach uses the 0.5 probability for all the prospects. They conclude that for people with high math skills, the preferred instrument is the HL as it fits the data better than the EG. However, for subjects with low math skills, they conclude that the EG device performs better both in terms of smaller noise and better fit. They base their analysis on Expected Utility Theory (EU) and use a Constant Relative Risk Aversion (CRRA) utility function combined with a random noise estimation that allows them to inspect how noise is associated with numeracy skills and other variables. Dave et al. (2010) also suggest that experimental devices that are designed so that they are easy to understand by persons with low math skills are more likely to find a real correlation between ability and risk attitudes.

Charness and Viceisza (2016) compare three risk preference elicitation approaches, the HL approach, a simple variant of the simple risky investment game (Gneezy, Leonard, & List, 2009) (GP) that builds on Gneezy and Potters (1997), and a non-incentivized willingness-to-take-risk (WTR) survey question, in a field experiment in rural Senegal with a focus on the performance of these three methods in such a setting. Although they did not measure numeracy skills we can assume that such skills are limited in their sample. They find that most respondents make inconsistent and dominant choices in the HL game even though it is framed in the local context. They therefore conclude

[1] The approach is similar to that of Binswanger (1980, 1981). While the EG approach involves a single choice among 6 gambles, the Binswanger approach includes dominated gambles, a non-linear trade-off between risk and return, and a pairwise comparison of prospects.

that using such a sophisticated mechanism is not effective in this type of rural environment. For the GP risky investment game, they find signs of arbitrary choices and Gillen, Snowberg, and Yariv (2019) find a high degree of randomness when assessing repeated decisions by students with high numeracy skills when using this game. In a more recent study, Charness, Eckel, Gneezy, and Kajackaite (2018) compare the standard HL approach with the use of only one row from the CL as a single binary decision. Their results indicate that subjects are better able to judge a single choice between two risky prospects than a list of ten with changing probabilities. It appears that the list is too complex and makes subjects confused.

Another approach that recently has been used more widely across the developed and developing country contexts is the Certainty Equivalent-Multiple Choice List (CE-MCL) approach (Holden & Tilahun, 2022; Vieider et al., 2018; Vieider, Martinsson, Nam, & Truong, 2019; Vieider et al., 2019). Some of the possible advantages of this approach to the HL is that there is only a single risky prospect in each CL and it is compared with alternative certain amounts without changing the probabilities in the risky prospect. Visual devices are used to illustrate the probabilities to subjects with limited numeracy skills. The approach has proved to have a high share of consistent responses. Vieider et al. (2018) find that 62% of a sample of rural respondents in Ethiopia make no between-CL inconsistent responses when responding to a set of 7 CLs. Holden and Tilahun (2022) find that 59% of their sample of rural respondents from Ethiopia gave no between-CL inconsistent responses when using 12 CLs and a rapid elicitation method to identify the switch point in each CL.

Our study builds on the studies by Holden and Tilahun (2022) and Vieider et al. (2018) and uses two variants of the CE-MCL design that to a varying degree can detect within-CL and between-CL decision errors[2] combined with a test of numeracy skills for a sample of 836 rural business group members in Ethiopia. We measured their numeracy skills with a 15-question test that was adapted to the business environment of our subjects. We inspect the extent of decision errors in the form of subject-level violations of stochastic dominance and how these are associated with the numeracy test score.

We aim to answer the following research questions: How capable are our study subjects to understand and respond rationally to our CE-MCL tools, given our step-wise elicitation approach, for the elicitation of risk preferences? Can we make designs that are simple enough for even those with very low numeracy skills to give rational responses that reveal their true preferences? And how much does variation in their limited numeracy skills contribute to decision errors and the estimated sizes of their risk preference parameters? Finally, we ask whether Expected Utility (EU) theory is sufficient or whether Rank Dependent Utility (RDU) does better in the analysis of decision errors and risk preferences in our context.

[2]Between-CL stochastic dominance is used to assess the consistency of responses at aggregate and subject levels.

We make a range of non-parametric tests to explore the nature of the relationship between numeracy skills, decision errors, and choice distributions in the two alternative CE-MCL designs. We find that the first CE-MCL design with 12 CLs per subject, which is close in design to that of Holden and Tilahun (2022), results in no between-CL consistency violations for 55% of the sample, based on 7 paired CL tests. The second CE-MCL design, consisting of 10 CLs per subject results in no consistency violations for 76% of a different sample from the same population, based on 4 paired CL tests. When the sample is split into four close to equally sized groups based on the number of correct answers in the numeracy test,[3] the average number of decision errors is only marginally higher for the group with the lowest numeracy skills score, indicating that also they can understand the binary questions in our CE-MCL approach the way it is introduced to them and give reasonably consistent answers.

We estimate structural models based on Expected Utility (EU) and Rank Dependent Utility (RDU)(Quiggin, 1982) theories while we separate decision errors.[4] We estimate the correlation between decision errors and numeracy skills in pooled and split sample models and find, as expected, that higher numeracy skills are associated with fewer decision errors. We also find that weaker numeracy skills are associated with slightly lower risk tolerance in the EU models, and with slightly stronger probabilistic insensitivity in the RDU models. The fact that the estimated Prelec α and β are significantly different from 1 implies that we should reject the EU model in favor of the RDU model.

The estimated average risk preference parameters for those with the lowest numeracy skills are only marginally different from the risk preferences of those with better numeracy skills and the number of decision errors is also only marginally larger, indicating that the elicitation approach is sufficiently simple for their intuitive numeracy skills to enable them to make reasonably rational decisions. This approach therefore appears well suited to be used among subjects with limited education and numeracy skills and should be preferred to the much-used HL approach which is more cognitively demanding to understand.

Our main contributions to the literature are the following. To our knowledge, this is the first comprehensive assessment of how numeracy skills are associated with decision errors and risk preferences in a developing country setting. Our study is the first to demonstrate that a fairly complex experimental tool can be presented simply to elicit dis-aggregated utility and probability weighting parameters from subjects with limited numeracy skills when it is made simple by splitting the comprehensive CE-MCL design into simple binary questions. Our study is the first to measure the extent to which weak and varying levels of numeracy skills affect decision errors and risk preferences in an EU versus an RDU framework and demonstrates that non-linear probability

[3]Numeracy skills are measured as a count variable and are therefore discrete. Therefore, the number of subjects per group is not exactly a quarter of the full sample.

[4]The RDU model is also in our setting consistent with the Cumulative Prospect Theory as we only study experiments in the gains domain(Tversky & Kahneman, 1992).

weighting with an inverse S-shaped w(p) function is a dominant characteristic. Our study has relevance for how to design field experiments to elicit risk preferences from populations with limited education.

Our paper proceeds as follows. Part 2 elaborates on the sampling, the orchestration of field experiments, and the numeracy skills test. Part 3 elaborates the experimental designs, and explores the experimental data quality with non-parametric methods and stochastic dominance assessment. Part 4 outlines the parametric estimation of structural models. Part 5 presents the structural model results, and Part 6 discusses the findings based on our main research questions and relevant literature before we conclude in Part 7.

2 Survey sample, Experimental Design, and Data

2.1 Sample characteristics

The business group program was established as a policy initiative to create a complementary natural resource-based livelihood opportunity for landless and near-landless youth and young adults in this risky environment (Holden & Tilahun, 2018, 2021). Eligibility criteria for joining the business groups were residence in the community and resource poverty in terms of limited land access. The main group production activities they could establish were animal rearing, beekeeping, forestry, and irrigation/horticulture. It enabled them to continue living in their home community close to their parents.

Basic socio-economic characteristics of our sample, by gender, are presented in Table 1.[5]

2.2 Sample and survey data

The study is based on a random sample of youth business groups from a census of 742 such groups in five districts in the semiarid Tigray Region of Ethiopia (Holden & Tilahun, 2018). Up to 12 members were sampled from each group, consisting of up to five group board members, and an additional random sample of ordinary group members. A baseline survey was implemented in July-August 2016. A second round of experiments and surveys were conducted in July-August 2017, and a third round of survey and experiments was implemented in July-August 2019. The groups and members included in each round changed from 2017 to 2019 (there is a limited overlap). This study is based on the 2019 round of the risk experiments. We used two alternative experimental designs and have a sample of 430 subjects for the first design (2019A) and 406 subjects for the second design (2019B). Table 1 provides some basic socioeconomic data for the sample.

[5]Unfortunately, we do not have years of education for the full sample used in this study. However, another study of the same sample population, based on a sample of 2400 subjects, found that the average number of years of education was 5.5 years (Holden & Tilahun, 2021).

Table 1 Basic sample characteristics (2019)

Variable	Males mean (sd)	Females mean (sd)
Sample size	451	385
Age	33.5 (8.97)	30.5 (7.35)
Married, dummy	0.80 (0.40)	0.80 (0.40)
Number of children	2.9 (2.17)	3.1 (1.97)
Group board member, dummy	0.46 (0.49)	0.22 (0.42)
Numeracy test score	4.5 (2.82)	3.7 (2.47)

2.3 Survey and risk experiment implementation

All experiments and survey questions were translated and asked in the local language, *Tigrinya*. Trained experimental and survey enumerators introduced the experiments and asked survey questions in the local language. Tablets and CSPro were the digital tools used for the data collection. Careful training of enumerators was first conducted in classrooms at Mekelle University. They were then trained by doing experiments and interviews with each other before they were trained in the field with out-of-sample groups and subjects. To minimize within-group spillover effects the twelve sampled members from each business group were interviewed simultaneously by 12 enumerators, using three classrooms in a local school or another local facility such as a Farm Training Centre. In schools, each enumerator was placed in the corner of each classroom and the subjects faced them during the experiments and survey interviews. Supervisors were used to ensure order and no disturbance. The orthogonal placement of enumerators on groups minimizes the risk of enumerator bias in the analyses. In addition, the researchers monitored potential enumerator bias during data collection and had follow-up meetings with the enumerators to identify reasons for observed enumerator bias in the data collected to find ways of minimizing such bias. Some poor-performing enumerators were replaced over the survey experimental rounds and others had to be replaced because they found other jobs. The enumerator team was stable within the 2019 survey and experimental round and they had been trained through participation in previous rounds.

2.4 Numeracy skills

Based on the 15 questions[6] basic math skills test, we constructed a simple score for the number of correct answers. The questions and the % correct responses to each question are presented in Appendix Table A1. The distribution of correct answers across the sample is shown in Figure 1. The average score (*numsum*) is as low as 4.1 correct answers out of 19. The median is 4 correct answers, p25=2, and p75=4 correct answers. The skewed distribution imposes some challenges in assessing the impact of (low) numeracy skills. We therefore

[6]Two of the questions required three answers, giving a maximum correct score of 19.

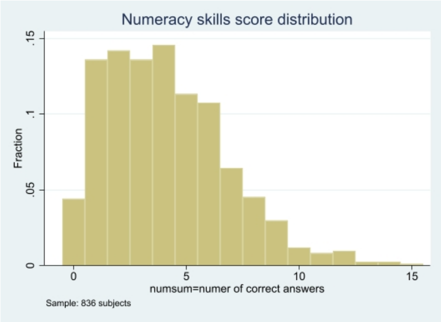

Fig. 1 The numeracy skills score distribution

use two different categorizations of numeracy skills. To capture substantial differences in numeracy skills we split the sample into those with very low numeracy skills with *numsum*$<$ 5, low numeracy skills with *numsum* in the range 5-9, and good numeracy skills for those with *numsum*$>$ 9. This gives the Categorization 1 distribution across the three classes of numeracy skills in Table 2. We see that the large majority of the subjects have very low or low numeracy skills. This gives reasons to worry whether the subjects at all are capable of understanding and responding in a rational way to our CE-MCL experiment (see next section). We hope that the use of simple binary choices, with illustrative devices (money on the table to demonstrate the risky prospect and alternative certain amounts, and the use of a 20-sided die to illustrate probabilities) can be understood even by subjects with so limited numeracy skills.

To further inspect the possible implications of the numeracy skills, we split the sample into four more even-sized groups by their numeracy skill score (Categorization 2 in Table 2 by *numsum* values 0-2, 3-4, 5-6, and >6.[7] The advantage of the first categorization is that it may help us to assess the effect of substantial differences in numeracy skills even though we have few observations at the high end of the skill score. The second categorization allows us to assess whether there are important differences between those with minimal numeracy skills and those with slightly better but low numeracy skills.

3 Experimental design

3.1 2019 CE-MCL Risk experiments

An almost identical experimental design to that used in 2017 by Holden and Tilahun (2022) was used in 2019 for 430 youth group members. Based on some concerns regarding this design,[8] a revised CL design was used for another

[7]The categorical distribution of *numsum* prevented us from splitting into four equal-sized groups.

[8]The certain amount range in the CLs was expanded for the CLs with higher probabilities of bad outcomes. This should potentially reduce the risk of censoring bias on the top of the lists. Such censoring was observed in the 2017 risk experiment and 2019A was also adjusted for two of the CLs that had this weakness in the 2017 design.

Table 2 Numeracy score categories 1 and 2

numsum	Freq.	Percent	Cum.
Categorization 1			
<5	506	60.5	60.5
5-9	301	36.0	96.4
>9	30	3.6	100.0
Total	837	100.0	
Categorization 2			
0-2	270	32.3	32.3
3-4	236	28.2	60.5
5-6	184	22.0	82.4
>6	147	17.6	100.0

sample of 405 youth group members. The first design, we call 2019A, focused particularly on a good mapping of the w(p) function in the more likely probability range for weather shocks in the form of droughts with p(drought) in the range 0.05 and 0.5 (Table 3). The new design, we call 2019B, has a balanced distribution of CLs across the p(bad outcome) range 0.05-0.95, with p-values 0.05, 0.2, 0.5, 0.8, 0.95 (Table 4). Both designs have risky prospects with bad outcomes at 0 and 20 ETB. Another difference is that we for the new design aimed to give each CL about the same expected value (about the same monetary incentive) by raising the good outcome amount as the probability of loss increases.

Some other noteworthy details are the following. CL 1 is identical in both the 2019A and 2019B designs. CL 12 in 2019A and CL10 in 2019B are identical except that the CE range goes from 30 to 120 (ev=75) in 2019A and goes from 30 to 300 in 2019B. 2019B thus opens for a larger degree of risk-loving behavior. This difference therefore facilitates a between-subject test for the effect of the CE-range expansion. CL 1 facilitates a test for whether the two sample distributions are different for p(bad)=0.95.

Both 2019A and 2019B facilitate several paired stochastic dominance comparisons of changing the bad outcome from 0 to 20 ETB. 2019A can in addition be used to test for stochastic dominance at aggregate and subject levels when the p(bad) is changing between 0.05, 0.1, and 0.2. CLs 11 and 12 in 2019A also have expected values (EV) close to each other (90 vs 94) and the same certain amount range and can be used to compare CE-equivalents for these CLs.

These experiments were implemented in July-August 2019 in combination with a follow-up survey of the same business groups and members. We used an elicitation approach where the subjects answered multiple series of binary questions where they in each question for a CL chose between a fixed risky prospect and an alternative certain amount. The advantage of this experiment is that it can separately identify the probability weighing function and the utility function, as we varied both probabilities and outcome levels (see Table 3 for an overview of key CL parameters). Table 4 provides an example of one of the CLs. The experimental protocol is included in the Appendix (Experimental Protocols).

The subjects are informed before the experiment is started that they will have to choose between a large number of risky prospects and certain amounts and that one of the prospects will be chosen randomly as a real game and for a real payout that will be given immediately after the experiment has been completed. Each subject is allocated to an MCL with a randomized order of the CLs. For each CL the subject is presented with the risky prospect which is outlined on the desk in front of her/him with real money for the good and bad outcomes and with the 20-sided die to illustrate the probability of winning and losing. It is only certain amounts that have to be changed to narrow in on the switch point and the CE for the risky prospect before the next CL and the risky prospect are outlined.

By holding the risky prospect constant, including the good and bad outcomes and the probability of good (bad) outcomes, we limit the required numeracy skills to deciding on the preferred choice between the risky prospect and the certain amounts.[9] Another advantage of this approach is that it is easy to present the risky prospect with real money in front of the subjects and illustrate the probabilities with the 20-sided die. In each CL a switch point is identified by the enumerator who uses a paper version of the CLs where the certain amounts are ordered in decreasing value from the top to the bottom of the CL. Tables 3 and 4 show the key characteristics of the 12 CLs used in the 2019A experiment and the 10 CLs in the 2019B experiment. The order of the CLs was randomized across subjects to allow assessment of and control for eventual order bias.

Concerns about starting point bias and bias towards the middle have been raised about the elicitation of preferences with CLs (Andersson, Holm, Tyran, & Wengström, 2016). Asking about every row from the top to the bottom and using many CLs per subject could also make subjects bored and create incentives to save time. We used an approach that aimed to address both these issues. To speed up the identification of the switch point in each CL, a quick narrowing-in approach was used. In each CL there is a randomized starting row number that identifies the certain amount that the risky prospect is to first be compared with. The quick elicitation approach means that the full CL is not presented to the subjects. The risky prospect is illustrated with real money in front of them with the probabilities demonstrated with the 20-sided die. The enumerators ask the subject to indicate their preference for the risky prospect or the certain amount at the random starting row in the CL as the first binary choice. The decision at this point identified whether the switch point would be above or below the random starting point certain amount. The enumerators were instructed to go to the top or the bottom of the list depending on the first choice. If subjects preferred the risky prospect at the random starting point, the CE-value of the risky prospect must be higher than the certain amount

[9] The well-known Holt and Laury (2002) is more demanding as it asks respondents to compare two risky prospects and at the same time changes the probabilities from row to row within the same CL and thereby demanding substantial numeracy skills and frequent recalculations.

at the starting row. The enumerator, therefore, goes to the top of the list and to the bottom row if the certain amount is preferred at the starting row. At the top of the list, we expect the respondents to prefer the certain amount.[10] Likewise, at the bottom of the list we expect respondents to prefer the risky prospect but here we added rows with lower certain amounts till a switch point was detected, meaning that the CE is below the lowest certain amount in the standard CL.[11] With a switch in the choice from the starting row to the top or bottom rows, a mid-row is chosen between the random starting row and the second (top or bottom row) in the CL, as the third decision row in the CL. Again the subject's choice in this third question is used to quickly narrow in towards the switch point as the two rows from where the subject switches from preferring the risky prospect to preferring the certain amount.

This bisection approach has several advantages; a) it reduces the number of questions per CL needed to identify the switch point (this reduces boredom and fatigue related to having to respond to many similar questions) and is therefore time-saving; b) the choices of random starting point reduces the likelihood of undetectable starting point bias such as if questions always start from one end of the CL; c) the potential bias associated with the random starting point can be tested and controlled for in the analysis;[12] d) a potential bias towards the middle of the CL is avoided as the whole list is not presented to the subjects;[13] e) the approach identifies only one switch point per CL (unless there is no switch point).

A context-specific design element of the CLs is that the risky prospect has two outcomes and the probability of a bad (but non-negative) outcome is stated to the subjects as a framing towards negative shocks. This framing is chosen as the experiment is intended used concerning behavior associated with low-probability shocks such as droughts. Droughts typically lead to low but non-negative yields.[14] Furthermore, 10 out of the 12 CLs in the 2019A design have prob(bad outcome)≤ 0.5, see Table 3. This also implies that we map most accurately the probability weighting function in the prob(bad outcome) range 0.05-0.5 with the 2019A design, the probability range within which most of the drought shocks may be found. The two last CLs in 2019A include a low probability of winning high return prospects to help us map the w(p) function also in this probability region. It is quite rare to have access to such business opportunities in our field context. Cultural norms and own experience may

[10] This may not always be the case and we then allow 'corner solutions' with CLs without any switch point. We return to the inspection of such outcomes and the remedies.

[11] We dropped two subjects with extreme risk aversion where we failed to detect a switch point as extremely small certain amounts were preferred to the risky prospects.

[12] This bisection approach has earlier been used in risk and time preference field experiments by Holden and Quiggin (2017a, 2017b); Holden and Tilahun (2022).

[13] Such bias has been an argument for placing the risk-neutral row at the center of the CL (Andersson et al., 2016) but would also lead to bias towards risk-neutrality for subjects that are risk averse.

[14] In Rank Dependent Utility (RDU) it is usual to sort outcomes from the best to the poorest (with their associated probabilities) and we do this in our structural model and estimation but we recognize that our framing gives higher salience to the negative shocks and this may have affected the responses in the intended way (focus on the non-negative bad outcomes and their probabilities).

therefore play less of a role in influencing their decisions in these CLs. This may cause a larger variance in the choices in these CLs.

In the end, the random choice of CL and Task row for payout is identified by the use of the 20-sided die using the underlying MCL. In the randomly identified CL for real payout, one task row is randomly identified and the subject's choice in this row determines whether the respondent will get the preferred certain amount or the preferred risky prospect. If the risky prospect was preferred for this row, the die is used to play the lottery and determine whether the subject receives a good or a bad outcome. The subject then received the outcome in cash in an envelope.

Table 3 Sample 2019A: CE-Multiple Choice List Treatment Overview

Choice List	Prob (bad outcome)	Bad outcome (ETB)	Good outcome (ETB)	CE-range min, max (ETB)
1	1/20	0	100	50,100
2	1/10	0	100	50,100
3	2/10	0	100	50,100
4	3/10	0	100	30,90
5	5/10	0	100	10,80
6	1/20	20	100	50,100
7	1/10	20	100	50,100
8	2/10	20	100	40,100
9	3/10	20	100	30,90
10	5/10	20	100	25,100
11	15/20	20	300	30,120
12	19/20	20	1500	30,120

Table 4 Sample 2019B: CE-Multiple Choice List Treatment Overview

Choice List	Prob (bad outcome)	Bad outcome (ETB)	Good outcome (ETB)	CE-range min, max (ETB)
1	1/20	0	100	50,100
2	2/10	0	150	50,150
3	5/10	0	250	70,200
4	8/10	0	600	70,200
5	19/20	0	1500	30,300
6	1/20	20	100	60,110
7	2/10	20	150	50,150
8	5/10	20	250	70,200
9	8/10	20	600	70,200
10	19/20	20	1500	30,300

Table 5 Example of Choice List, CL 8 in 2019A

CL no.	Start point	Task no.	Prob. low outcome	Low outcome	High outcome	Choice	Certain amount	Choice
8		1	2/10	20	100		100	
8		2	2/10	20	100		95	
8		3	2/10	20	100		90	
8		4	2/10	20	100		85	
8		5	2/10	20	100		80	
8		6	2/10	20	100		75	
8		7	2/10	20	100		70	
8		8	2/10	20	100		65	
8		9	2/10	20	100		60	
8		10	2/10	20	100		50	

3.2 Experimental outcome distributions and data quality

The cumulative switch point distributions in the 2019A risk CE-MCL experiment are presented in Fig. 2-4, with CLs 1-3 and CLs 6-7 in Fig. 2. The stochastic dominance is very clear from the graphs demonstrating that the CE falls with an increasing probability of a bad outcome in Fig. 2a and 2b. The combined CLs in Fig. 3a and 3b only differ in the size of the bad outcome, also with very clear stochastic dominance effects. It is also noteworthy for CL1 and CL6 that the risk-neutral row is row 2 (or very close to row 2 for CL6).[15] For this low probability of a bad outcome, close to 90% of the subjects are risk averse and prefer the certain amount. For CL2 and CL7 the risk-neutral row is row 3 or just below (for CL7) where about 90% of the subjects are risk averse and switch for CE<EV.

Fig.4, the second graph, shows the cumulative distributions for CL11 and CL12 (low probability (0.25 and 0.05) high outcomes (ETB 300 and 1500)). The higher shares of corner solutions without switch points in CL11 and CL12 indicate a higher willingness to take the risk for such low probability high outcomes.[16] Only about 80% have CE<EV for these CLs.

We further investigate the balance of the CLs in terms of the location of the means and medians in terms of switch point rows in each CL for the 2019A and 2019B designs. With a standard of 10 rows per CL, a balanced switch point pattern should have the mean and median switch point rows at rows 5 to 6 in each CL. Tables 6 and 7 show the mean and median rows just above the switch points in the 2019A and 2019B designs. For 2019A in Table 6 we see that the median row varies from 3 (CL 6) to 8 (CL 3) with an overall median of 6. For 2019B in Table 7 we see that median switch point rows above the switch points vary from 5 (CLs 1 and 6) to 8 (CLs 3, 4, 5, and 10) and with an overall median at 7. This indicates a stronger skewness toward the bottom of the CLs for 2019B than for 2019A. Our approach of adding rows at

[15] The certain amount offered is 95 in this row.

[16] With hindsight we see that we should have included even higher certain amounts at the top of these CLs.

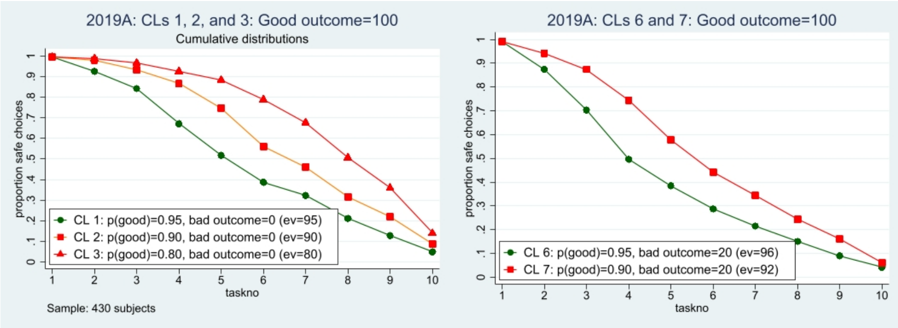

Fig. 2 2019A: The distribution of switch points in CL 1-CL 3 and CL 6-CL 7

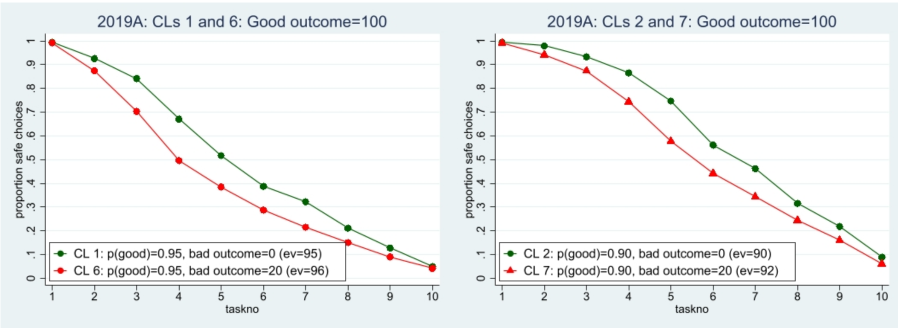

Fig. 3 2019A: The distribution of switch points in CL 1 vs. CL 6 and CL 2 vs. CL 7

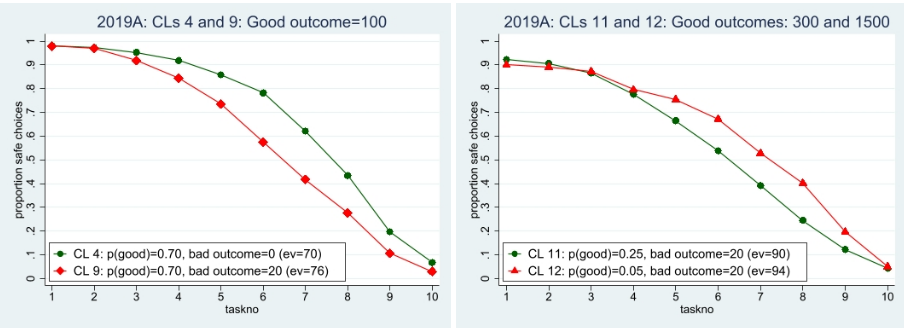

Fig. 4 2019A: The distribution of switch points in CL 4 vs. CL 9 and CL 11 vs. CL 12

the bottom in the cases when subjects had not switched to the risky prospect at row 10 prevented censoring at the bottom. The fact that we did not use a similar procedure at the top of the CLs implies that we should more carefully inspect for possible censoring at this end of the CLs. About 8 to 10% of the sample for CLs 11 and 12 in 2019A may suffer from this problem (Fig. 4b) even though the median switch point rows were 6 and 7 for these CLs and the EV-rows were 5 and 4.4 (CE<EV).

Fig. 5 2019B: The distribution of switch points in CL 2 vs. CL 7 and CL 3 vs. CL 8

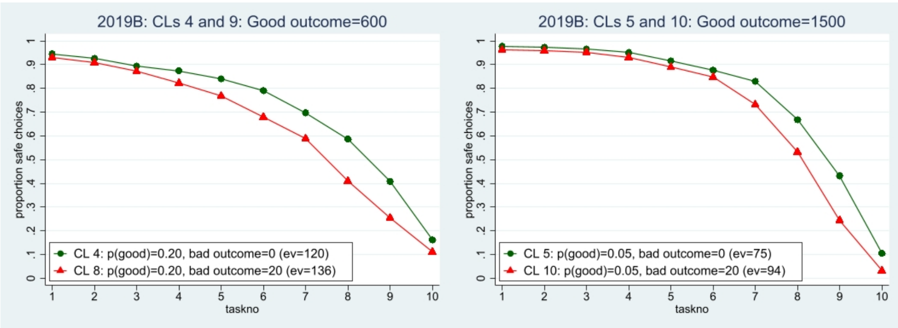

Fig. 6 2019B: The distribution of switch points in CL 4 vs. CL 9 and CL 5 vs. CL 10

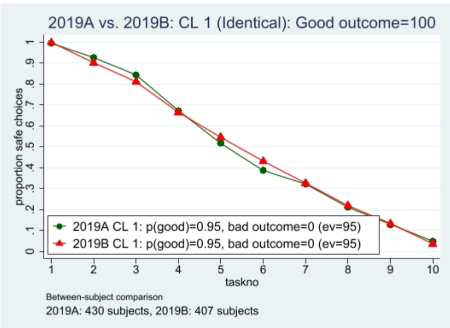

Fig. 7 Test of similarity of 2019A and 2019B: CL1 Identical across samples

Fig. 8 and 9 present the cumulative switch point distributions by the row number just above the switch points for all the CLs in 2019A and 2019B. Here we inspect in particular for censoring at the top of the lists. Such censoring could be due to decision errors associated with random choice at the top of the list.[17] If this kind of random choice error is common, such decision errors

[17]Recall that respondents that prefer the risky prospect in the first question for a CL at the randomly chosen row, will be asked about the choice at the top row in the following question. If the subject still prefers the risky prospect, we have a censored observation.

Table 6 2019A: Mean and median switch point rows vs EV-rows in each CLs

CL	Mean	Median	EV-row	St.err.	N
1	5.1	5	2	0.041	4329
2	6.3	6	3	0.039	4354
3	7.4	8	5	0.036	4379
4	6.8	7	4	0.034	4329
5	5.8	6	4	0.033	4306
6	4.3	3	1.8	0.042	4329
7	5.5	5	2.6	0.041	4339
8	5.5	5	2.6	0.038	4328
9	5.9	6	2.8	0.035	4313
10	5.8	6	5	0.028	4315
11	5.5	6	5	0.041	4317
12	6.1	7	4.4	0.044	4318
Total	5.8	6		0.011	51956

Table 7 2018B: Mean and median switch point rows vs EV-rows in each CLs

CL	Mean	Median	EV-row	St.err.	N
1	5.1	5	2	0.042	4073
2	6.7	7	3.5	0.035	4092
3	7.7	8	4.67	0.037	4149
4	7.3	8	5	0.046	4151
5	7.8	8	8.5	0.033	4098
6	5.0	5	2.8	0.039	4081
7	5.7	6	3.2	0.036	4077
8	6.7	7	4.25	0.039	4104
9	6.5	7	4.2	0.046	4117
10	7.1	8	7.87	0.034	4065
Total	6.6	7		0.013	41007

should occur in all CLs. However, we do not find that and this is a clear indication that such random choice at the top of the lists is not a problem. For 2019A in Figure 8, we observe censoring at the top of the lists only for two CLs, in CLs 11 and 12 (the initial step at row 0 for the row just above the switch point in the cumulative distributions indicates censoring and no switch point). This is a sign that for low p(win) probability CLs, a (small) share of the sample is more willing to take risks and therefore needs to be given higher certain amounts than those at the top of these CLs to identify their CEs. In 2019B we included higher certain amounts at the top of the CLs for the low p(win) CLs (CLs 4, 5, 9, and 10). Fig. 9 shows a longer left-side tail for these CLs but even here we have not eliminated all top-censoring indicating that some behave as strong risk-lovers in these CLs.[18]

To further inspect the data quality we inspect for stochastic dominance violations at the subject level as another way to detect decision errors. The

[18]We can also easily read out the median switch point row from these graphs and the extent to which we had to add extra rows below row 10 to identify the switch point for part of the two samples.

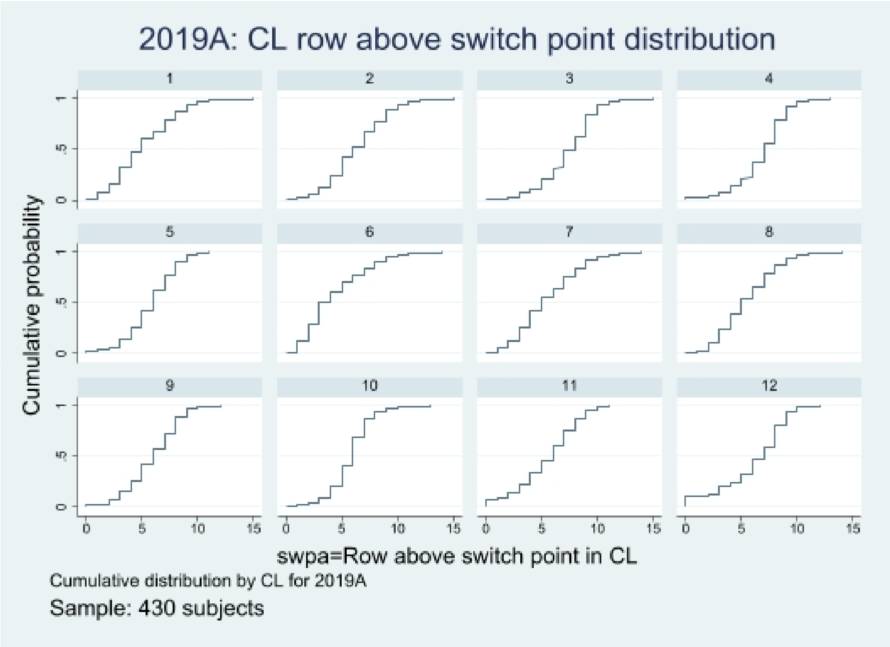

Fig. 8 Switch point row cumulative distributions by CL in 2019A

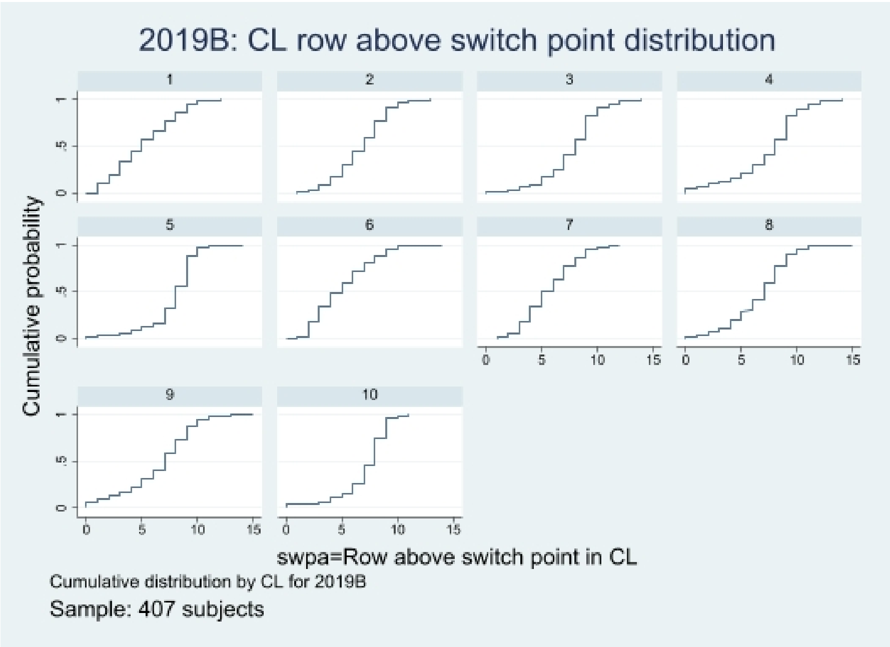

Fig. 9 Switch point row cumulative distributions by CL in 2019B

two CL-MCL designs have different powers in detecting such decision errors. 2019A facilitates the detection of errors for the size of the bad outcome as well as for relatively small differences in the probabilities of bad (and good) outcomes, while 2019B only can be used to detect the first type of decision error. For stochastic dominance assessment, two CLs must have all except one parameter that are the same across the (within-subject) CLs.[19] All in all, we can make 7 paired subject-level stochastic dominance comparisons with 2019A and 4 with 2019B. The extent of violations by CL pairs is presented in Table 8 and the distribution of the number of violations per subject for 2019A and 2019B are presented in Table 9.

We see that the 2019A design was more likely to detect decision errors and this may also affect the designs' ability to assess how numeracy skills limit the ability to separate decision errors from risk preferences.

When we look at the aggregated distribution of subject-level across-CL stochastic dominance violations in our 2019A sample based on the assessment above (seven paired comparisons per subject), we find that 55.0% had no violations, 22.1% had one violation, 17.5% had two violations, 4.2% had three violations, and 1.6% had four violations in the case of the 2019A design. The 2019B design gave fewer within-subject across-CL stochastic dominance violations with 76.1% having no violations, 17.5% having one violation, 5.7% with two violations, and 0.7% with three violations.

Our ability to assess how numeracy skills affect decision errors and separate decision errors from risk preferences depends on the experimental designs and 'the devil is in the details' of the designs. 2019A is more capable than 2019B of detecting decision errors in the form of stochastic dominance violations. We can assess how the subject-level number of stochastic dominance violations is associated with the numeracy test scores. Table 10 gives an overview of the average number of violations by numeracy score Category 2 separation into groups 0-2, 3-4, 5-6, and above 6 correct answers for the 2019A and 2019B designs. The table shows that there is no significant reduction across the four numeracy test score groups in average number of violations for the 2019A design, while for the 2019B design the group with $numsum > 6$ has a significantly lower average number of violations (mean=0.153, se=0.047) vs. means 0.282, se=0.054 for the closest of the other groups ($numsum 3-4$) among those with lower numeracy scores.

4 Theoretical Framework and Estimation

To further investigate the extent of correlation between numeracy skills and risk preferences, we estimate structural models based on EU and RDU theories where we separate heteroskedastic noise with a Fechner error based on Wilcox contextual utility (Wilcox, 2008, 2011). This specification allows us to assess how decision errors are related to numeracy skills and CL design

[19]The number of such paired comparisons that can be made differs for the 2019A and 2019B designs as we illustrated in the aggregated stochastic dominance assessment in Figures 2-4 for 2019A and Figures 5-6 for 2019B.

Table 8 Subject level stochastic dominance violations (%) for paired CLs

	Observations	Percent
Choice List 2019A:		
CL1-CL2	429	12.4
CL2-CL3	429	12.8
CL1-CL3	429	5.4
CL6-CL1	429	10.5
CL7-CL2	429	12.8
CL9-CL4	429	9.8
CL6-CL7	429	11.9
Choice List 2019B:		
CL7-CL2	405	9.4
CL8-CL3	405	6.9
CL9-CL4	405	5.9
CL10-CL5	405	8.9

Table 9 Subject level number of violations of stochastic dominance

	Frequency	Percent	Cumulative
Choice List 2019A:			
No violations	236	55.0	55.0
One violation	95	22.1	77.2
Two violations	73	17.5	94.2
Three violations	18	4.2	98.4
Four violations	7	1.6	100
N	429	100	
Choice List 2019B:			
No violations	308	76.1	76.1
One violation	71	17.5	93.6
Two violations	23	5.7	99.3
Three violations	3	0.7	100
N	405	100	

Based on 7 and 4 paired CL comparisons (Table 7)

characteristics while at the same time assessing whether numeracy skills are related to the structural model parameters. We estimate pooled models for the 2019A and 2019B designs as well as separate models for each design. Detailed specifications of the structural parametric models follow.

4.1 EU and RDU model estimation

Each choice of the subject is between a risky prospect and a certain amount. The risky prospect gives a good outcome (x) with probability p and a bad outcome (y) with probability $1 - p$. We call the certain amount s. We place the choice between the risky and safe prospect alternatively into an Expected Utility (EU) or a Rank Dependent Utility (RDU) framework (Quiggin, 1982). The net utility return for a specific risky and safe option can then be formulated

Table 10 Number of stochastic dominance violations in 2019A and 2019B by numeracy score category

	2019A			2019B		
numsum	mean	se(mean)	N	mean	se(mean)	N
0-2	0.769	0.084	130	0.396	0.055	139
3-4	0.838	0.100	117	0.282	0.054	117
5-6	0.720	0.090	107	0.351	0.082	77
>6	0.640	0.113	75	0.153	0.047	72
Total	0.753	0.048	429	0.311	0.030	405

as follows:

$$\Delta RDU = w(p)u(x) + [1 - w(p)]u(y) - u(s) \tag{1}$$

where $w(p)$ is the probability weighting function. The more general RDU model nests the EU model where $w(p) = p$. In a specific CL, x and y are fixed while s varies across the rows with falling values from the top. There will be a point where the ΔRDU switches from being negative (preference for larger certain amounts s), to becoming positive (preference for the risky prospect over smaller certain amounts s). The certainty equivalent (CE) is identified at the switch point as the average value of the certain amounts in the rows just below and above the switch point.

The CE-MCL risk experiment included only prospects with non-negative outcomes.[20] The probability weighting function is therefore modeled in the gains domain only with a Prelec et al. (1998) 2-parameter weighting function:

$$w(p) = e^{-\beta(-\ln p)^{\alpha}}, \alpha > 0, \beta > 0 \tag{2}$$

where α captures the degree of (inverse) S-shape of the weighting function[21], and the β captures the elevation of the function, with $\beta < 1$ giving more elevated (optimistic) and $\beta > 1$ giving less elevated (pessimistic) weighting of prospects. The function is strictly increasing and continuous within the interval $[0, 1]$ with $w(0) = 0$ and $w(1) = 1$. Probabilistic insensitivity at medium probability levels is the most common finding in studies of probability weighting, implying an inverse S-shaped probability weighting function(Prelec et al., 1998; Vieider et al., 2019).

[20] There are ethical reasons for not introducing experiments with negative outcomes to the type of poor and vulnerable subjects that are the focus of this study. However, the 2019A format was also related to a drought context where the bad outcomes could be zero or a very low non-negative output level. Another reason for us avoiding going into a Prospect Theory (PT) approach is that reference points are unobservable, and may be subject-specific, making it extremely difficult to comprehensively implement PT with endogenous reference points in this kind of setting. This would require the identification of subject-level reference points, loss aversion, and w(p) functions separately in the gains and loss domains. It was clearly beyond what we were able to do in this already very challenging field context given the low levels of numeracy skills of the research subjects.

[21] $\alpha = 1$ implies $w(p) = p$, for $\alpha < 1$ the inverse S-shape becomes stronger as α declines

The utility is modeled with a Constant Relative Risk Aversion (CRRA) function[22]

$$u(x) = (1-r)^{-1}((b+x)^{1-r} - 1) \tag{3}$$

where r is the Constant Relative Risk Aversion (CRRA) coefficient and b is the base consumption or asset integration level.[23]

Noise in the data is captured with a heteroscedastic Fechner (1860) type error (ξ) and the prospects are standardized with Wilcox (2008) type contextual utility.[24]

Contextual heteroscedasticity can be due to the error variance increasing with the subjective utility ranges. Wilcox (2008) argues that the contextual utility model uses the idea that the standard deviation of evaluation noise is proportional to the subjective range of stimuli, borrowing from the perception of stimuli literature, e.g. Gravetter and Lockhead (1973). This implies the assumption that each CL creates its own respondent-specific 'local context'.

The probability of the respondent choosing the risky lottery can then be formulated with a probit (standard normal) function:

$$Pr(Risky) = \phi(\frac{\Delta RDU_{gimk}}{\xi_{gimk}[u(x_m) - u(y_m)]}) \tag{4}$$

As explained before subscripts g, i, m, and k represent groups, subjects, CLs, and row numbers in the CLs. The model flexibility allows respondent errors in the identification of switch points within CLs. The latent Fechner error (ξ_{gim}) which is multiplied with the CL-specific utility difference between the good and bad outcome for the risky prospect, can be assessed at the within-subject CL level as a measure of subject response inconsistency across CLs or at a higher structural model level to assess model performance.

The log-likelihood function (equation 5) for the risk experiment is obtained by summing the natural logs over the cumulative density functions resulting from equation 4 and summing them over CLs (subscript m) and subjects:

$$\begin{aligned} &\ln L(\Omega_{gi}(numsumD_{gi}, mcl_{gi}), \xi_{gimk}(numsumD_{gi}, mcl_{gi}, cl_m, E_d)) = \\ &\sum_{imk} (\ln \Theta(\Delta RDU) \mid_{Choice_{imk}=1}) + (\ln \Theta(1 - \Delta RDU) \mid_{Choice_{imk}=0}) \end{aligned} \tag{5}$$

[22]We assume incomplete (no or partial) asset integration based on the finding that prospect amounts have much stronger effects on decisions than the respondents' background wealth (Binswanger, 1981).

[23]We set the base consumption equal to 0 ETB in the models in this paper (assume no asset integration).

[24]According to Wilcox (2008) the advantage of this approach is that the assessment of choices fits within the theoretical idea of capturing stochastically more risk-averse behavior without introducing extra parameters. Wilcox also states that binary choice models are better at measuring ratios of utility differences than utility differences. Utility differences need to be judged within their specific context. This is a fundamental problem in this kind of structural latent variable discrete choice model. Utilities have to be judged against a salient utility difference. Wilcox suggests using the utilities of the maximum and minimum possible outcomes in the risky prospect. This implies that choices are directly weighted by the subjective range of utility outcomes while holding marginal utility improvements constant near a maximum (Wilcox, 2008).

Ω_{gi} is a vector of subject-specific risk preference parameters $(r_{gi}, \alpha_{gi}, \beta_{gi})$ that only is allowed to vary with the Category 1 three-level numeracy test score or the Category 2 four-level numeracy test score $(numsumD)$[25], represented by two alternative dummy vectors. mcl_{gi} represents the random order of CLs that vary across subjects and groups and cl_m represents CL-specific fixed and random characteristics.

$$\Omega_{gi} = \eta_0 + \eta_D numsumD_{gi} + \epsilon_{gi} \tag{6}$$

The Fechner error (ξ_{gimk}) is also modeled as a function of the three-level numeracy skills variable.

$$\xi_{gimk} = \rho_1 + \rho_{2D} numsumD_{gi} + u_{gimk} \tag{7}$$

4.2 Decision error analysis

We are interested in investigating how numeracy skills are associated with decision errors, or possibly risk preferences more directly. We use the structural model with Fechner error to separate decision errors from risk preference estimation. We investigate the Fechner error sensitivity to the Category 1 numeracy skill dummy vector, the random starting row within the CL (CL_{str}), the within-subject random order of the CL (CL_{ro}) (learning and boredom effects), the CL-specific placement of the risk-neutral row within the CL (CL_{rn}), and the placement of the random starting row versus the risk-neutral row within the CL (CL_{str-rn}). We do this by including controls for these in the Fechner error specification (equation (8). Finally, we included a vector of enumerator dummy variables (E_d) for the random enumerators allocated to each subject within each business group.[26]

$$\begin{aligned}\xi_{gimk} = \rho_{1e} + \rho_{2eD} numsumD_{gi} + \rho_{3e} CL_{str} + \rho_{4e} CL_{ro} \\ + \rho_{5e} CL_{rn} + \rho_{6e} CL_{str-rn} + \rho_{7e} E_d + u_{gimk}\end{aligned} \tag{8}$$

With this error specification, we estimate pooled models and models for the two different CE-MCL designs 2019A and 2019B. This gives a robustness test for the alternative designs for our key research questions.

Next, we do sample-splitting by Category 2 numeracy skill level in four groups with close to similar sample sizes per group to assess whether there are significant differences in risk preference parameters and Fechner errors across numeracy skill levels (without and with the additional within-CL controls in equation (8)). This allowed us to assess whether the additional decision error controls were correlated with the numeracy skill level of the subjects.

We estimated the likelihood function with the Newton-Raphson optimization algorithm while clustering errors at the subject level.

[25]Alternatively, by the inclusion of additional subject characteristics, more of the sample variation in the risk preferences can be traced.

[26]The ability of enumerators to minimize respondent errors may vary. 12 enumerators were randomly allocated to subjects within groups.

Table 11 Pooled sample EU and RDU models with 3-level numeracy skills (Category 1)

EQUATION	VARIABLES	(1) EU model	(2) RDU model
CRRA-r	1.numsumD	-0.005	0.056*
		(0.026)	(0.033)
	2.numsumD	-0.095**	0.023
		(0.046)	(0.048)
	Constant	0.492***	0.185***
		(0.017)	(0.024)
Prelec α	1.numsumD		0.039**
			(0.017)
	2.numsumD		0.086***
			(0.033)
	Constant	1.000	0.617***
		(0.000)	(0.010)
Prelec β	1.numsumD		-0.075
			(0.050)
	2.numsumD		-0.129**
			(0.064)
	Constant	1.000	1.290***
		(0.000)	(0.036)
Fechner error	1.numsumD	-0.031**	-0.014
		(0.014)	(0.011)
	2.numsumD	-0.084***	-0.059***
		(0.016)	(0.010)
	Constant	0.235***	0.168***
		(0.010)	(0.008)
	Observations	92,963	92,963

Models with Wilcox contextual utility. NumsumD=0 as base category.
Cluster-robust standard errors in parentheses, clustering on subjects.
Significance levels: *** p<0.01, ** p<0.05, * p<0.1

5 Results

5.1 Pooled EU and RDU models testing for numeracy skills effects

Table 11 presents the parsimonious structural EU and RDU models where the 3-level numeracy skills variable (Category 1) is included as two dummy variables with the lowest numeracy skills group as the base category. It gives us a first test of whether numeracy skills only are correlated with the Fechner error or also is significantly correlated with the utility curvature and the Prelec w(p) parameters in the RDU model.

Both models show significant correlations between the Fechner error and numeracy skills and with a stronger reduction in the Fechner error with higher numeracy skills in the EU model than in the RDU model, where we only find a significant reduction in error for the highest numeracy skills group. The EU model gives a more concave utility function than the RDU model and the concavity is significantly lower (more risk tolerant) for the group with the highest numeracy skills, unlike in the RDU model where the results point weakly in the opposite direction. In the RDU model, the numeracy skill effect or correlation is strong in the Prelec α equation which demonstrates a significantly stronger inverse S-shape for the group with very low numeracy skills with α=0.617, increasing to 0.656 for the intermediate group, and to 0.703 for the group with the highest numeracy skills. This indicates that probabilistic insensitivity is associated with weak numeracy skills. When it comes to the Prelec β parameter, we see from Table 11 that those with low and very low numeracy skills are quite pessimistic while those with the highest numeracy skills are a bit less pessimistic.

5.2 EU and RDU models split by CE-MCL design and including Fechner error CL controls

Table 12 compares a pooled sample EU model with Fechner error CL controls with split-sample models for the two CE-MCL designs we have abbreviated as the 2019A and 2019B designs. The key results for numeracy skills are similar in direction and sign for the CRRA-r parameter as in Table 11 while the numeracy effects on the Fechner error are lower in size and significance. On the other hand, the CL design variables in the Fechner error equations were significant and appear to have controlled for some of the numeracy skill effects. The CL order variable is significant and with a negative sign in all three models, indicating that there has been some learning that contributed to reducing the error over multiple CLs. A higher random start row number (further down in the CL) is associated with a higher error in the pooled and the 2019B models. The placement of the risk-neutral row further down in the CL is associated with smaller errors in the split models but not the pooled model. The placement of the random starting row relative to the risk-neutral row below the risk-neutral row is associated with higher error. However, the effect of these additional controls in the Fechner error equation appears not to have had a strong effect on the estimated CRRA-r parameter with only a slight reduction from 0.49 to 0.47 in the pooled model without and with these additional controls.

The results for the pooled and split-sample RDU models with additional controls in the Fechner equations are presented in Table 13. The results for the Prelec α parameter appear robust concerning the direction of the numeracy skill effect on the parameter and the size of the parameter (0.59-0.64 for the group with weakest numeracy skills and 0.69-0-75 for those with the strongest numeracy skills). The effect of numeracy skills on pessimism bias was less robust, being insignificant in the 2019B design and the pooled sample after the inclusion of the additional error controls.

When it comes to the utility function, it is slightly concave for those with the weakest numeracy skills in the pooled sample and for the 2019B sample while it is close to linear for those with better numeracy skills in the 2019B sample. Overall, the utility curvature is close to linear with CRRA-r in the range -0.03-0.18 across all these models.

Table 12 EU models: Numeracy score, decision errors, and CE-MCL design

EQUATION	VARIABLES	(1) Full sample	(2) 2019A	(3) 2019B
CRRA-r	1.numsumD	-0.017	-0.028	-0.020
		(0.023)	(0.028)	(0.032)
	2.numsumD	-0.092**	-0.104**	-0.111*
		(0.045)	(0.051)	(0.064)
	Constant	0.469***	0.482***	0.443***
		(0.016)	(0.019)	(0.023)
Fechner error	CL order	-0.006***	-0.004***	-0.010***
		(0.001)	(0.001)	(0.002)
	1.numsumD	-0.012	0.001	-0.023*
		(0.009)	(0.008)	(0.013)
	2.numsumD	-0.044***	-0.028**	-0.061***
		(0.014)	(0.012)	(0.021)
	Start row in CL	0.002**	0.001	0.003***
		(0.001)	(0.001)	(0.001)
	Risk-neutral row in CL	0.001	-0.013***	-0.004*
		(0.002)	(0.002)	(0.002)
	Start row - Risk-neutral row	0.008***	0.006**	0.010***
		(0.001)	(0.003)	(0.001)
	Constant	0.292***	0.321***	0.362***
		(0.024)	(0.028)	(0.034)
	Observations	92,763	51,956	40,807

Models with Wilcox contextual utility. NumsumD=0 as base category (Category 1 grouping). Cluster-robust standard errors in parentheses, clustering on subjects. Significance levels: *** $p<0.01$, ** $p<0.05$, * $p<0.1$

The additional controls in the Fechner error equation also reduced the size of the numeracy skills effects but they remained significant and negative for the highest numeracy skill category in the pooled and the split 2019A and 2019B designs. While the CL order effect was robust and in the same direction as in the EU models, the results for the other controls deviated from those in the EU models in some interesting ways. The starting row number was insignificant in the 2019B design but was highly significant in the 2019A design. The placement of the risk-neutral row had a significant effect in opposite directions in the 2019A and the 2019B designs. It goes beyond the scope of this paper to dive into the possible explanations for these differences.

5.3 EU and RDU-models with sample split by numeracy skills

As a further robustness check of the importance of numeracy skills for the estimation of the structural parameters in the EU and RDU models, we split the sample by numeracy skill level in four close to equal sample sizes (Category

Table 13 RDU models: Numeracy score, decision errors, and risk preference parameters

EQUATION	VARIABLES	(1) Full sample	(2) 2019A	(3) 2019B
CRRA-r	1.numsumD	0.026	0.054	0.013
		(0.035)	(0.044)	(0.032)
	2.numsumD	-0.060	0.138	-0.160**
		(0.055)	(0.089)	(0.080)
	Constant	0.130***	-0.032	0.169***
		(0.022)	(0.035)	(0.020)
Prelec α	1.numsumD	0.034**	0.019	0.041**
		(0.016)	(0.022)	(0.018)
	2.numsumD	0.087**	0.107*	0.102***
		(0.035)	(0.057)	(0.036)
	Constant	0.633***	0.641***	0.586***
		(0.011)	(0.016)	(0.011)
Prelec β	1.numsumD	-0.053	-0.089*	-0.037
		(0.045)	(0.051)	(0.053)
	2.numsumD	-0.028	-0.215**	0.082
		(0.068)	(0.084)	(0.126)
	Constant	1.322***	1.449***	1.302***
		(0.031)	(0.037)	(0.038)
Fechner error	CL order	-0.004***	-0.002**	-0.003***
		(0.001)	(0.001)	(0.001)
	1.numsumD	0.003	0.001	-0.004
		(0.007)	(0.008)	(0.007)
	2.numsumD	-0.027***	-0.037***	-0.023**
		(0.009)	(0.011)	(0.009)
	Start row in CL	0.002***	0.004***	0.001
		(0.001)	(0.001)	(0.001)
	Risk-neutral row in CL	-0.014***	0.005***	-0.018***
		(0.002)	(0.002)	(0.001)
	Start row - Risk-neutral row	-0.000	-0.002	-0.001**
		(0.001)	(0.002)	(0.001)
	Constant	0.250***	0.175***	0.233***
		(0.019)	(0.021)	(0.020)
	Observations	92,763	51,956	40,807

Models with Wilcox contextual utility. NumsumD=0 as base category (Category 1).
Cluster-robust standard errors in parentheses, clustering on subjects.
Significance levels: *** p<0.01, ** p<0.05, * p<0.1

2 grouping).[27] The EU models are presented in Table 14 and the RDU models in Table 15.

Table 14 shows that the Fechner error has a declining trend with increasing numeracy skills as we would expect. The difference is not very large though. We also see a weak tendency of declining average CRRA-r with increasing numeracy skills but the differences across groups are not significant.

[27] Category 1 grouping of numeracy skills is less suitable for sample splitting because of the uneven sample sizes per group.

Table 14 EU models by Category 2 numeracy skill level

EQUATION	*numsum−− >*	(1) 0-2	(2) 3-4	(3) 5-6	(4) >6
CRRA-r	Constant	0.497***	0.485***	0.478***	0.477***
		(0.023)	(0.024)	(0.026)	(0.025)
Fechner error	Constant	0.237***	0.233***	0.204***	0.192***
		(0.015)	(0.014)	(0.013)	(0.012)
	Observations	29,680	26,055	20,709	16,519

Models with Wilcox contextual utility. Sample split by Category 2 numeracy skill score. Cluster-robust standard errors in parentheses, clustering on subjects. Significance levels: *** p<0.01, ** p<0.05, * p<0.1

Table 15 RDU models by Category 2 numeracy skill level

EQUATION	*numsum−− >*	(1) 0-2	(2) 3-4	(3) 5-6	(4) >6
CRRA-r	Constant	0.204***	0.163***	0.229***	0.247***
		(0.033)	(0.034)	(0.031)	(0.029)
Prelec α	Constant	0.615***	0.620***	0.659***	0.662***
		(0.014)	(0.015)	(0.017)	(0.018)
Prelec β	Constant	1.273***	1.310***	1.212***	1.207***
		(0.051)	(0.051)	(0.044)	(0.045)
Fechner error	Constant	0.170***	0.165***	0.157***	0.142***
		(0.012)	(0.010)	(0.010)	(0.010)
	Observations	29,680	26,055	20,709	16,519

Models with Wilcox contextual utility. Sample split by Category 2 numeracy skill score. Cluster-robust standard errors in parentheses, clustering on subjects. Significance levels: *** p<0.01, ** p<0.05, * p<0.1

The RDU models in Table 15 show a weak but significant increase in the Prelec α parameter as the numeracy skills score increases but all average parameters stay in the narrow range of 0.615-0.662, indicating very stable results overall. The Prelec β is in the 'pessimistic' range 1.21-1.31 for all numeracy levels. The Fechner error is falling slightly with the numeracy skill score. The sample splitting by numeracy skill score therefore revealed results consistent with the previous findings for the pooled models. There are only small and barely significant differences across numeracy skill groups for the utility curvature and Prelec β parameters. The effects of numeracy skills on the Fechner error and the Prelec α parameter are significant but the magnitude differences are not large. The most interesting result, given our first research question, is that even those with the weakest numeracy skills appear to have understood the simple CE-MCL approach with binary questions quite well and have not performed so poorly that the results need to be discarded.

Table 16 EU models split by numeracy skill level, with additional Fechner error controls

EQUATION	*numsum*−− >	(1) 0-2	(2) 3-4	(3) 5-6	(4) >6
CRRA-r	Constant	0.467***	0.444***	0.441***	0.454***
		(0.022)	(0.024)	(0.022)	(0.026)
Fechner error	CL order	-0.005***	-0.009***	-0.005**	-0.003*
		(0.002)	(0.003)	(0.002)	(0.002)
	Start row in CL	0.002	0.002*	0.002	0.000
		(0.001)	(0.001)	(0.001)	(0.001)
	Risk-neutral row in CL	-0.006**	0.000	0.003	0.004
		(0.003)	(0.003)	(0.003)	(0.002)
	Start row - Risk-neutral row	0.013***	0.007***	0.008***	0.010***
		(0.003)	(0.002)	(0.002)	(0.003)
	Constant	0.308***	0.352***	0.248***	0.231***
		(0.038)	(0.046)	(0.044)	(0.042)
	Observations	29,580	26,055	20,709	16,419

Models with Wilcox contextual utility. Sample split by Category 2 numeracy skill score. Cluster-robust standard errors in parentheses, clustering on subjects. Significance levels: *** p<0.01, ** p<0.05, * p<0.1

Finally, we have estimated models with the sample-splitting by numeracy skill level combined with the expanded Fechner error controls. The results are presented in Tables 16 and 17 for the EU and the RDU models. The EU models in Table 16 reveal no significant differences in utility curvature across numeracy skill levels. Compared to the EU models in Table 14 without the additional Fechner controls, the addition of the Fechner controls resulted in a slightly less concave utility function for all the numeracy skill levels. A similar tendency is found for the utility curvature in the RDU models when comparing Tables 15 and 17. The results for the w(p) parameters were very robust. The CL order effects were also robust across models. The additional Fechner error controls gave more mixed results and went beyond what we aim to investigate and discuss in this paper.

6 Discussion

6.1 Discussion of key research questions

We start the discussion by first addressing our key research questions. Our first research question is 'How capable are businessmen and women with very limited education of understanding and responding rationally to economists' tools for the elicitation of risk preferences?' Several studies that have used the HL approach have found that people with limited numeracy skills have problems responding consistently to this approach (Charness & Viceisza, 2016; Dave et al., 2010). Two studies that used a row-by-row approach with the HL design found that people are more capable of understanding that than giving rational responses when presented with whole lists (Charness et al.,

Table 17 RDU models split by numeracy skill level, with additional Fechner error controls

EQUATION	*numsum*−− >	(1) 0-2	(2) 3-4	(3) 5-6	(4) >6
CRRA-r	Constant	0.090***	0.081***	0.171***	0.143***
		(0.030)	(0.030)	(0.033)	(0.032)
Prelec alpha	Constant	0.625***	0.642***	0.696***	0.667***
		(0.014)	(0.015)	(0.018)	(0.019)
Prelec beta	Constant	1.381***	1.363***	1.231***	1.284***
		(0.045)	(0.045)	(0.038)	(0.047)
Fechner error	CL order	-0.005***	-0.006***	-0.003*	-0.004***
		(0.001)	(0.002)	(0.002)	(0.002)
	Start row in CL	0.002	0.004***	0.002*	-0.000
		(0.001)	(0.001)	(0.001)	(0.002)
	Risk-neutral row in CL	-0.021***	-0.015***	-0.007*	-0.016***
		(0.002)	(0.002)	(0.004)	(0.002)
	Start row - Risk-neutral row	-0.000	-0.001	0.001	0.002
		(0.001)	(0.002)	(0.001)	(0.001)
	Constant	0.279***	0.288***	0.211***	0.285***
		(0.026)	(0.031)	(0.034)	(0.046)
	Observations	29,580	26,055	20,709	16,419

Models with Wilcox contextual utility. Sample split by Category 2 numeracy skill score. Cluster-robust standard errors in parentheses, clustering on subjects. Significance levels: *** p<0.01, ** p<0.05, * p<0.1

2018; Holden & Quiggin, 2017b). However, the standard change in probabilities from row to row in the HL-CL for two risky prospects demands substantial numeracy skills and is more cognitively demanding than our CE-MCL design where there is only one risky prospect in each CL and with a constant within-CL probability of good and bad outcomes. It is this feature that makes our approach more attractive for the elicitation of risk preferences of subjects with limited numeracy skills. We used a 20-sided die to illustrate the probabilities of good and bad outcomes in each CL.

Our elicitation technique facilitated rapid identification of the switch point in each CL. It prevented multiple switch points in each CL and we have primarily assessed subject-level decision errors across CLs. A crucial finding with our method is that random choice is not a big problem. The fact that all subjects are asked about their decision in either the top row or the bottom row as the second question for each CL (depending on their decision in the first randomly chosen row), gives us the opportunity to assess whether random choice is common. With a high risk of random choice in the second question when this is at the top of the lists, we would get many censored responses there. We do not find that. Only for the low p(win) CLs did we get some censoring at the top. We can therefore rule out that random choice represents a problem in the binary questions, and this is the case for those with lowest numeracy skills as well.

Additional subject-level within-CL decision errors can only be inspected at the aggregate level with the additional controls in the extended Fechner error specification. For the within-subject across-CL decision errors, we find a high share of the subjects making no such decision errors and the average number of decision errors was not very high even for those with the weakest numeracy skills. The estimated average EU and RDU parameters only responded marginally to the variation in numeracy skills.

The other added advantage of our approach is that it allows us to test the EU model against the RDU model. Simpler approaches and the HL approach are unsuited or less well suited for this. Overall, we find that our subjects with limited numeracy skills are better modeled with the RDU approach than the EU approach as $\alpha = 1$ and $\beta = 1$ can be rejected and the Fechner error is also smaller in the RDU models than in the EU models (e.g. see Table 14 vs. Table 15), showing that the former structural model explains more of the variation in the data. Non-linear probability weighting in the form of an inverse S-shaped w(p) function is a fundamental characteristic of our study subjects. Our sample is located in the pessimistic range for the probability weighting function with Prelec β values in the range 1.2-1.5.

Based on this assessment, we are now ready to also answer our second research question: 'Can we make designs that are simple enough for those with very weak or weak numeracy skills such as a large share of our study sample to give rational responses that reveal their true preferences?'

We have exposed our respondents to a variety of CL designs and can reveal the extent to which they have responded consistently when exposed to 12 or 10 CLs each. As they were not exposed to the whole CL at the time but only made binary choices when faced with one row at the time it is the pattern of binary choices that helps us to assess their capacity to respond rationally. Our results reveal a high share of rational responses even among those with the lowest numeracy skills. With an average of only 0.77 stochastic dominance violations out of 7 in the 2019A design and 0.40 out of 4 in the 2019B designs the group with the lowest numeracy skills performed astonishingly well. It shows us that they have a clear understanding of risky versus certain amounts of money.

We have demonstrated that our CE-MCL design has the potential to be so simple that even subjects with very limited numeracy skills, such as our sample, can understand and respond well to the binary questions with our randomized rapid elicitation approach. It is both simple and sophisticated and is better suited for the elicitation of RDU parameters than the HL design because we hold the probabilities constant within CLs and vary them systematically across CLs. Our rapid elicitation approach facilitates the use of more CLs per subject without exhausting them. The fact that the decision error declines with the random order of the within-subject CLs shows that learning dominates over boredom and inattention when we have used 10 and 12 CLs per subject.

Our final research question is 'And how much does variation in their limited numeracy skills contribute to decision errors and the estimated sizes of their risk preference parameters?' Most of our subjects had very low or low numeracy skills and we have only 3.8% of the sample having more than 50% correct responses in the contextualized numeracy test. The Fechner error is significantly smaller, about 35% smaller in both the pooled EU model and the pooled RDU model for this group compared with the group with the poorest numeracy skills (Table 11). When it comes to the effects of numeracy skills on the risk preference parameters, we also compare the small group with more than 50% correct answers with the group with less than 25% correct answers (60% of the sample). Table 10 shows that the first group has a significantly lower CRRA-r in the EU model (0.40 vs. 0.49) while this significant difference is not there in the RDU model, which instead finds significant differences in the Prelec α (0.70 vs. 0.62) and Prelec β (1.16 vs. 1.29) parameters. This illustrates that the choice of very simple tools and only EU theory as the basis for eliciting risk preferences may fail to capture important characteristics of the risk preferences of subjects with limited numeracy skills. The RDU theory combined with the CE-MCL design with binary choices appears more suited and it appears to work reasonably well even for subjects with very limited numeracy skills.

Our alternative sample splitting approach of dividing the sample into four more equally sized groups based on their numeracy test scores provides further evidence that even the group with only 0-2 correct responses on the numeracy test gave reasonably few average number of decision errors (Table 9) and average parameter estimates close to those with better numeracy test scores.

6.2 A comparison with other studies

We will now compare our findings with other studies. We did not find any other studies that directly measured numeracy skills and risk preferences in a developing country setting. However, there is substantial literature on cognitive ability and how it may be related to risk preferences (Dohmen et al., 2018; Lilleholt, 2019; Mechera-Ostrovsky et al., 2022). There is also relevant literature on financial literacy and economic outcomes (Hastings, Madrian, & Skimmyhorn, 2013; Lusardi & Mitchell, 2014). Financial literacy or numeracy skills are a type of cognitive skills that also have been characterized as crystallized intelligence in the psychology literature (Cattell, 1971, 1987). There is ample evidence that such skills are important for economic success and welfare. However, financial skills are endogenous and can be related to underlying latent cognitive ability. Our study does not aim to test the effect of numeracy skills on risk preferences or economic outcomes. We only aim to investigate whether we can estimate latent risk preferences given the limited numeracy skills in our sample. It thus goes beyond our scope to test for causality between risk preferences and numeracy skills. We would need some exogenous treatment that influences numeracy skills to assess their potential impact on risk preferences. E.g. we cannot conclude whether the numeracy skills are driving the revealed risk preferences or whether they only influence perceptions that lead to wrong interpretations of the risk experiments and thereby e.g. the probabilities. However, we tentatively argue that the results indicate that low numeracy skills may contribute to stronger probabilistic insensitivity in the intermediate range of probabilities. We find it less plausible that probabilistic risk preferences of this nature cause low numeracy skills. We leave it open whether one should regard probabilistic insensitivity as a decision error or a specific type of probabilistic risk preference. Probabilistic insensitivity is widespread also among subjects with higher education and strong numeracy skills.

There are a couple of recent systematic reviews and meta-analyses that investigate the relationship between cognitive ability and risk preferences. Lilleholt (2019) reviewed 97 studies in the gains domain, 41 in the mixed domain, and 12 in the loss domain. This review found a weak but significant positive relationship between cognitive ability and risk tolerance in the gains domain but no significant correlations in the mixed and loss domains. This is consistent with our finding with the EU model in pooled models and models with the two variants of the CE-MCL design. Changing to RDU models, this result was no longer consistent in our data if the focus was only on risk aversion measured with utility curvature. Lilleholt (2019) did not test such alternative structural models with probability weighting, thus it is not possible to draw any conclusions from this meta-study about the underlying structural relations. Another limitation is that it was not possible to control for decision errors and this may cause a downward bias in the correlation coefficient between cognitive ability and risk tolerance as suggested by the author.

Another recent meta-study assessing whether cognitive ability affects risk preferences concludes that it does not (Mechera-Ostrovsky et al., 2022). They conclude that cognitive abilities affect decision errors and such decision errors may cause spurious correlations between risk preferences and cognitive ability, depending on the design of the elicitation device. They tested an alternative 'error hypothesis' which states that lower cognitive abilities increase decision errors against the 'risk preference hypothesis' which states that cognitive abilities affect the evaluation of choice options and thereby risk-taking behavior. They carried out a systematic review and meta-analysis based on 30 studies. Their study revealed no credible association between cognitive abilities and risk tolerance. They conclude that apparent correlations between cognitive ability and risk preferences were due to biases in the risk preference estimates caused by decision errors.

The extent of bias due to spurious correlations between random choice errors and risk preferences depends on the design of e.g. CLs. The further towards the top or the bottom of a CL the true switch point lies, the more likely it is that a random choice around the true switch point hits the top or bottom row, thereby censoring the distribution. Such censoring can lead to biased estimates. Andersson et al. (2016) argued that the risk-neutral row should be placed in the middle of a CL to avoid bias such that subjects appear as more risk averse than they are. We argue that if one does not know whether subjects are risk-neutral on average, the mean or median subject should have the switch point at the middle of a CL. However, the extent of censoring at the corners may also depend on the size of the tails in distributions that may not necessarily be symmetric as we saw in some of our CLs in Figures 8 and 9. Especially for low p(win) for the risky prospect, we observed long left-side tails, indicating that some individuals appear very risk-loving or less cautious in such CLs. We found mean and median switch points at row numbers below the risk-neutral row in all but two of our CLs, see Tables 6 and 7. Only for CLs 5 and 10 in Table 7 do we find mean and median switch point rows in the distributions to be slightly above or very close to the EV-row row number.[28] Our approach of adding rows at the bottom of the CL in cases when subjects still preferred the certain amount at the bottom row, also prevented censoring at this end of the lists. In combination with the separation of subject decision errors in the estimation, we therefore think we have been able to go far in reducing bias due to random choice.

Not only cognitive abilities but also cognitive attention and motivation can reduce decision errors and possibly lead to biases in estimated preference parameters. Our study revealed a reduction in the Fechner error associated with the random order of the CLs that the subjects responded to. This is likely due to subject-level learning as they repeatedly respond to the binary questions used to identify their switch points in each CL. The rapid elicitation approach has reduced the time required for each CL and therefore also the time costs of using such a large number of CLs per subject. Our study shows that the Fechner error is reduced by the cognitive skills, measured with our numeracy test score, but our experimental device appears to have performed well even for those with the lowest numeracy score as their Fechner error is only about 20% higher than that of the group with the highest numeracy test score (Tables 14 and 15).

Hey, Morone, and Schmidt (2009) studied noise and bias related to four different methods for the elicitation of risk preferences. These included pair-wise choices among two risky prospects, maximum buying price for lotteries, minimum selling price for lotteries, and the reporting of certainty equivalent for lotteries (based on the Becker-DeGroot-Marschak mechanism). They had a small sample of 24 student subjects that were in the register of the Centre of Experimental Economics and the University of York so their subjects lived in a very different context from our context and were obviously at a very different level of numeracy skills. Nevertheless, there may be some generalized insights of relevance to our study as well. For the pair-wise choice between lotteries, these were presented with amounts and probabilities on a computer screen. The question asked related to the last certainty equivalent experiment was 'State the amount of money such that you do not care whether you will receive this amount or the lottery'. From this, it is obvious that our CE-MCL with binary choices is closer to the pair-wise choices in their experiments than their certainty equivalent elicitation approach, and other experimental approaches. Their generalized statement that certainty equivalent experiments

[28]We note that the mean and median rows in Tables 5 and 6 are the rows immediately above the switch point while the EV-row is placed exactly at the switch point, implying that the gap between the mean and median versus the EV-rows is even larger in most cases.

lead to biased results does therefore not apply to our study. On the contrary, their finding that pair-wise choice was associated with the smallest decision errors supports our pair-wise (binary) approach for the elicitation of CE switch points. Unlike, what we have done with a much larger sample, Hey et al. estimated models for the four methods for each subject separately. This allowed them to do within-subject comparisons across methods. An interesting finding in their study was that their single probability weighting parameter in the RDU model was more strongly correlated across methods than the CRRA-r was, indicating more stable attitudes towards probabilities than outcomes. However, they found signs of only weak probabilistic insensitivity compared to what we found and this may be due to the much stronger numeracy skills of the subjects in their sample. Another interesting finding was that they found the pair-wise comparison method to result in significantly more concave utility both with the EU and the RDU specifications. This is in contrast to our study where the EU models give more concave utility functions than the RDU models.

There is a misconception that the CE-MCL approach needs to lead to a certainty bias. Vieider (2018) showed that such a bias can be turned in opposite direction through a modification of the design. By presenting the risky prospect in each CL and keeping it constant and asking binary questions about the preference for the risky prospect and one alternative certain amount at the time, the risky prospect is more salient and this should not lead to a certainty bias.

7 Conclusions

Our study is to our knowledge the first to assess the relationship between numeracy skills, the extent of decision errors, and how these correlated with risk preferences in a developing country setting. Based on a separate numeracy test adapted to the local business environment and two variants of a Certainty Equivalent-Multiple Choice List design, with a binary elicitation procedure, we tested for the extent of within-subject decision errors through stochastic dominance assessments. Additional decision error tests associated with CL design, random orders of CLs, and random starting rows in each CL in structural EU and RDU models with a separate heteroscedastic Fechner error specification, were used in combination with tests for numeracy skills scores in pooled and split sample models. Weaker numeracy skills were associated with lower risk tolerance (more concave utility) and larger Fechner errors in EU models. In RDU models weaker numeracy skills were associated with less probabilistic sensitivity and larger Fechner errors. However, even the subjects with the lowest numeracy skills (<12% correct responses) had an average Fechner error that was only 20% larger than the group with highest numeracy skills (>35% correct responses) and the average number of stochastic dominance violations in the two CE-MCL designs was only moderately larger than for those with the better numeracy skills.

Non-linear probability weighting with an inverse S-shaped w(p) function is a stable characteristic for the whole sample with Prelec in the range $0.6 - 0.7$ with those with lowest numeracy skills in the lower end of this range and those with better numeracy skills in the higher end of this range. Prelec β was significantly higher than one and in the range $1.2 - 1.4$ indicating somewhat pessimistic expectations that explain risk aversion more than the local utility function which was found to be close to linear with CRRA in the range $0.08 - 0.17$ in the RDU models split by numeracy skill score.

We conclude that the CE-MCL with a binary elicitation procedure is capable of eliciting the risk preferences of subjects with limited education and numeracy skills as long as careful procedures are followed and the subjects are motivated to answer the questions. We did not detect any problem with random choice causing censoring at the top of our CLs. The approach has advantages over other simpler procedures that are less able to control for decision errors and unable to handle non-linear probabilistic sensitivity along a probability scale. Estimation based on Rank Dependent Utility appears preferable to estimation based on Expected Utility.

Appendix A Numeracy skills questions and test results

See separate file.

Appendix B Experimental protocol

See separate file.

Supplementary information. Experimental designs are attached in a separate pdf-file.

Acknowledgments. This research has been conducted as a collaboration between the Norwegian University of Life Sciences (NMBU) and Mekelle University. The authors acknowledge good support from local government authorities, local Youth Associations, and Mekelle University, and committed efforts by our team of enumerators and field supervisors. Valuable comments have been received on an earlier draft from Dag Einar Sommervoll.

Declarations

- Funding
  Data collection: Stein T. Holden and Mesfin Tilahun, Grant Number ETH-13/0015 Norwegian Agency for Development Cooperation (NORAD), the NORHED I capacity building project "Climate Smart Natural Resources Management and Policy"

(https://www.norad.no/en/front/funding/norhed/projects/capacity-building-for-climate-smart-natural-resource-management-and-policy–clisnarp–/) and own research fund of the first author. Data cleaning, organization and analyses: Stein T. Holden Grant Number: 288238 The Research Council of Norway https://www.forskningsradet.no/en/. The funding institutions had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

- Conflict of interest/Competing interests
  The authors declare no conflicts of interest.
- Ethics approval
  Funding was approved based on an independent assessment and approval of ethical standards being met by the project by a scientific committee.
- Consent to participate
  All subjects participating in the project participated voluntarily and were always asked up-front about their willingness to participate after having received information about what participation implied and that the project adhered to strict confidentiality and anonymity of individual information (informed consent).
- Consent for publication
  The article will be published as an open-access article as required by the funding institution.
- Availability of data and materials
  All (anonymized) data (STATA files) used in the paper will be made available upon publication of the article as supplementary information.
- Code availability
  All codes (Stata do files) used for the analysis of the data will be made available upon publication as supplementary files.
- Authors' contributions
  The first author made the initial experimental designs. Both authors collaborated on the field testing of the survey and experimental designs and the training of enumerators. The second author was in charge of all the data collection and organizing survey and experimental teams. The first author was in charge of data checking and both contributed to data cleaning and organization. The first author wrote up the paper and the second author commented on the drafts.

References

Andersson, O., Holm, H.J., Tyran, J.-R., Wengström, E. (2016). Risk aversion relates to cognitive ability: Preferences or noise? *Journal of the European Economic Association*, *14*(5), 1129–1154.

Binswanger, H.P. (1980). Attitudes toward risk: Experimental measurement in rural india. *American Journal of Agricultural Economics*, *62*(3), 395–407.

Binswanger, H.P. (1981). Attitudes toward risk: Theoretical implications of an experiment in rural india. *Economic Journal*, *91*(364), 867–890.

Cattell, R.B. (1971). Abilities: Their structure, growth, and action.

Cattell, R.B. (1987). *Intelligence: Its structure, growth and action.* Elsevier.

Charness, G., Eckel, C., Gneezy, U., Kajackaite, A. (2018). Complexity in risk elicitation may affect the conclusions: A demonstration using gender differences. *Journal of risk and uncertainty*, *56*, 1–17.

Charness, G., & Viceisza, A. (2016). Three risk-elicitation methods in the field-evidence from rural senegal. *Review of Behavioral Economics*, *3*(2), 145–171.

Cokely, E.T., Feltz, A., Ghazal, S., Allan, J., Petrova, D., Garcia-Retamero, R. (2018). Decision making skill: From intelligence to numeracy and expertise. *Cambridge handbook of expertise and expert performance*, 476–505.

Dave, C., Eckel, C.C., Johnson, C.A., Rojas, C. (2010). Eliciting risk preferences: When is simple better? *Journal of Risk and Uncertainty*, *41*(3), 219–243.

Dohmen, T., Falk, A., Huffman, D., Sunde, U. (2018). On the relationship between cognitive ability and risk preference. *Journal of Economic Perspectives*, *32*(2), 115–34.

Eckel, C.C., & Grossman, P.J. (2002). Sex differences and statistical stereotyping in attitudes toward financial risk. *Evolution and human behavior*, *23*(4), 281–295.

Eckel, C.C., & Grossman, P.J. (2008). Forecasting risk attitudes: An experimental study using actual and forecast gamble choices. *Journal of Economic Behavior & Organization*, *68*(1), 1–17.

Fechner, G.T. (1860). *Elemente der psychophysik* (Vol. 2). Breitkopf u. Härtel.

Garcia-Retamero, R., Sobkow, A., Petrova, D., Garrido, D., Traczyk, J. (2019). Numeracy and risk literacy: What have we learned so far? *The Spanish Journal of Psychology*, *22*, E10.

Gillen, B., Snowberg, E., Yariv, L. (2019). Experimenting with measurement error: Techniques with applications to the caltech cohort study. *Journal of Political Economy*, *127*(4), 1826–1863.

Gneezy, U., Leonard, K.L., List, J.A. (2009). Gender differences in competition: Evidence from a matrilineal and a patriarchal society. *Econometrica*, *77*(5), 1637–1664.

Gneezy, U., & Potters, J. (1997). An experiment on risk taking and evaluation periods. *The Quarterly Journal of Economics*, *112*(2), 631–645.

Gravetter, F., & Lockhead, G. (1973). Criterial range as a frame of reference for stimulus judgment. *Psychological Review*, *80*(3), 203.

Hastings, J.S., Madrian, B.C., Skimmyhorn, W.L. (2013). Financial literacy, financial education, and economic outcomes. *Annu. Rev. Econ.*, *5*(1), 347–373.

Hey, J.D., Morone, A., Schmidt, U. (2009). Noise and bias in eliciting preferences. *Journal of Risk and Uncertainty*, *39*, 213–235.

Holden, S.T., & Quiggin, J. (2017a). Bounded awareness and anomalies in intertemporal choice: Zooming in google earth as both metaphor and model. *Journal of Risk and Uncertainty*, *54*(1), 15–35.

Holden, S.T., & Quiggin, J. (2017b). Climate risk and state-contingent technology adoption: shocks, drought tolerance and preferences. *European Review of Agricultural Economics*, *44*(2), 285–308.

Holden, S.T., & Tilahun, M. (2018). The importance of ostrom's design principles: Youth group performance in northern ethiopia. *World Development*, *104*, 10–30.

Holden, S.T., & Tilahun, M. (2021). Preferences, trust, and performance in youth business groups. *Plos one*, *16*(9), e0257637.

Holden, S.T., & Tilahun, M. (2022). Are risk preferences explaining gender differences in investment behavior? *Journal of Behavioral and Experimental Economics*, 101949.

Holt, C.A., & Laury, S.K. (2002). Risk aversion and incentive effects. *American Economic Review*, *92*(5), 1644–1655.

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

Lusardi, A., & Mitchell, O.S. (2014). The economic importance of financial literacy: Theory and evidence. *American Economic Journal: Journal of Economic Literature*, *52*(1), 5–44.

Mechera-Ostrovsky, T., Heinke, S., Andraszewicz, S., Rieskamp, J. (2022). Cognitive abilities affect decision errors but not risk preferences: A meta-analysis. *Psychonomic Bulletin & Review*, *29*(5), 1719–1750.

Prelec, D., et al. (1998). The probability weighting function. *Econometrica*, *66*, 497–528.

Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior & Organization*, *3*(4), 323–343.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323.

Vieider, F.M. (2018). Violence and risk preference: experimental evidence from afghanistan: comment. *American Economic Review*, *108*(8), 2366–82.

Vieider, F.M., Beyene, A., Bluffstone, R., Dissanayake, S., Gebreegziabher, Z., Martinsson, P., Mekonnen, A. (2018). Measuring risk preferences in rural ethiopia. *Economic Development and Cultural Change*, *66*(3), 417–446.

Vieider, F.M., Martinsson, P., Nam, P.K., Truong, N. (2019). Risk preferences and development revisited. *Theory and Decision*, *86*(1), 1–21.

Vieider, F.M., et al. (2019). All over the map: A worldwide comparison of risk preferences. *Quantitative Economics*, *10*(1), 185–215.

Wilcox, N.T. (2008). Stochastic models for binary discrete choice under risk: A critical primer and econometric comparison. *Risk Aversion in Experiments*, *12*, 197–292.

Wilcox, N.T. (2011). 'stochastically more risk averse:'a contextual theory of stochastic discrete choice under risk. *Journal of Econometrics*, *162*(1), 89–104.

Zhang, H., & Holden, S.T. (2023). Numeracy skills learning of children in africa:—are disabled children lagging behind? *Plos one*, *18*(4), e0284821.