Too hard to get: the role of probabilistic expectations and cognitive complexity in multi-dimensional reference dependence

Aspen Han

This paper seeks to investigate the effects of conflicting reference points across different dimensions of utility on effort exertion. Reference-dependent preference models so far have assumed additive separability across different dimensions of utility, which implies that agents respond to reference points in each dimension in isolation from one another. Challenging this assumption, this paper posits that agents consider multi-dimensional reference points in tandem: agents are less responsive to reference points if they have low probabilistic expectations of being able to concurrently achieve them and/or if they have difficulty reconciling them into a single baseline against which to evaluate outcomes. It refines the Koszegi-Rabin reference-dependent preference model to incorporate these effects and applies it to examine effort exertion under targets in different task performance dimensions. The original and refined model produce distinct predictions for optimal effort exertion, which are tested via a real effort experiment.

Introduction

Consider the employee of a firm whose performance is evaluated against targets across various performance dimensions (e.g. production speed, accuracy, quality etc). For example, an assembly line worker in an electronics manufacturing plant could be subject to targets on the number of components made per hour (speed), the proportion of defective components made (accuracy), and the average durability of components made (quality). Similarly in the service sector, an Uber driver could be evaluated on the number of rides provided per month, average mileage per unit time, and the average customer satisfaction rating. It is apparent that tensions between these performance dimensions can surface, which can affect the targets' effectiveness as motivators. The emphasis for consistency and complementarity between different performance dimensions, including targets set in each, is strongly echoed in operations and general management literature (e.g. Hayes 1984; Hayes and Schmenner 1978; Skinner 1974, 1996; Swamidass and Darlow 2000). I seek to examine this concept within economics. Targets can and have been integrated into the framework of expectations-based reference points (Heath, Larrick, and Wu 1999; Von Rechenberg, Gutt, and Kundisch 2016), a growing body of research within behavioral economics. However, empirical studies have mainly examined the effects of reference points uni-dimensionally, though theoretical models encompassing multi-dimensional reference points exist. Thus, I wish to investigate the mechanisms through which reference points interact across dimensions within this economic framework, specifically answering the following research questions:

1. Do probabilistic beliefs about the achievability of reference points across multiple dimensions affect how responsive agents are to said reference points?
2. Does cognitive complexity in reconciling reference points across multiple dimensions affect how responsive agents are to said reference points?

This research is theoretically founded on the Koszegi and Rabin (2006) model of reference-dependence (henceforth KR model), which builds upon Kahneman and Tversky's (1979; 1991) prospect theory and related models of regret and disappointment (insert citations). Defining features of this overarching framework include the evaluation of outcomes relative to a reference point rather than on absolute terms, weighting losses more than gains, diminishing sensitivity away from the reference point, and decision weights on outcomes which are transformations of objective probabilities. A major contribution by the KR model is that it constructs a source for the reference points to be the agent's (rational) expectations, specifically his/her probabilistic beliefs held in the recent past about what will or should happen. This accommodated alternative arguments about the origins of reference points, such as the status quo (e.g. Genesove and Mayer 2001; Kahneman, Knetsch, and Thaler 1990) and refutes to it (e.g. Plott and Zeiler 2005; Tversky and Kahneman 1991). Pinpointing the source of reference points allowed for more detailed studies into their manipulability from a policy perspective, which motivates my use of the KR model as a theoretical baseline. The model also generalized to multiple dimensions of consumption, increasing portability to the conventional consumer choice problems in economics. While there remains continued debate about the source of reference points, this paper abstracts away from that, instead focusing on whether the KR model is a good description of behavior with respect to reference points across multiple dimensions[1] within the framework of expectations-based reference point. Similar to other expected utility and reference-dependent models, the KR model assumes that utilities across different dimensions of consumption are additively separate, which I seek to challenge. Given the motivating example highlighted above, it seems unrealistic to think that people would view reference points in each dimension in isolation from one another and determine how much to work towards each with complete disregard for the others.

In fact, it was precisely a contradiction to the additive separability assumption from my undergraduate research findings which inspired this study. The interaction effects between reference points

[1] The KR model also has other descriptively contentious components, such as modelling utility of a realized outcome to be the sum of both neoclassical consumption utility (absolute outcome levels) and gain-loss utilities (relative outcome levels), and the probability weighting function to be the identity of objective probabilities, though these do strengthen the model's normative appeals.

across dimensions had been largely neglected despite a rich empirical literature in the expectations-based reference points. Crawford and Meng (2011) found that the work patterns of New York taxi drivers could be explained by the KR model with dual reference points in daily wages earned and hours worked. While this is one of few works to consider multi-dimensional reference points, the field context made it difficult to elucidate the reference points, much less the mechanisms through which they could have interacted and affected the drivers' work behavior. Furthermore, since the taxi drivers are independent contractors, their reference points are self-imposed and hence likely consistent by construction, whereas conflicting effects are the focus of my research questions. Abeler et al. (2011) tested and verified the KR model in a laboratory experiment where subjects were set reference points in earnings and then asked to work on a real effort task. The controlled setting allowed the reference points to be exogenously induced so their effects on effort provision could be explicated. However, they only considered a reference point in a single dimension and hence neglected multi-dimensional interaction effects. Thus, synthesizing the laboratory methodology of Abeler et al. and the dual reference point model of Crawford and Meng my undergraduate research sought to test the multi-dimensional version of the KR model. It found that when the two reference points were congruent, they had reinforcing effects, which fits with KR model predictions, but when they were conflicting, they had negating effects in that subjects seemed to ignore the reference points completely instead of compromising between them or prioritizing one over the other as predicted by the KR model. This leads to my research questions, which endeavor to identify the reasons behind this destructive effect between disparate reference points in different dimensions.

I propose two main explanations: agents are unresponsive to reference points when they perceive the probability of being able to achieve them concurrently to be low, and/or when they find it cognitively complex to reconcile the reference points, and these problems arise when reference points across multiple dimensions conflict. We can easily append these features to the KR model through additional parameters which scale the gain-loss utility components, which would alter the first-order conditions predicting optimal effort provision such that they align with the experimental results.

I tested these propositions with a laboratory experiment. I elected for ~~a~~ experimental methodology as I wanted to clearly identify the decision-making mechanisms which integrate multi-dimensional reference points, and this is most clearly elicited in the controlled experiments, being difficult to establish with observational data where the reference points are elusive and there are many potential confounds. While I have linked my research motivations to the workplace, the foremost step would be to uncover the general ways in which people perceive and respond to multi-dimensional reference points which are applicable to various contexts, so the abstract setting of the laboratory experiments is well-suited for it. It also provides a less costly way to verify the hypothesized mechanisms at work, so it can be thought of as a pilot for further field research.

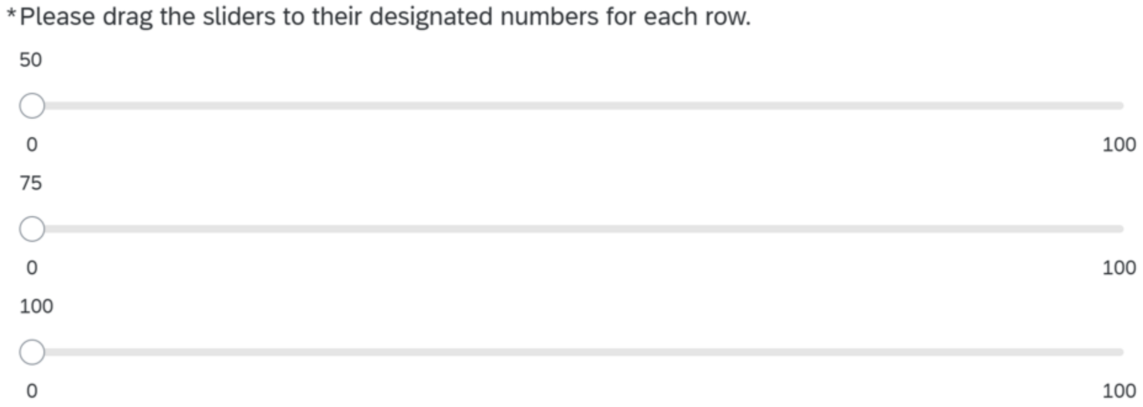
In the experiment, subjects worked on a real effort task where they have to drag sliders along a scale of 0 to 100 to designated numbers. They were evaluated on speed as measured by the number of slider sets completed per minute and accuracy as measured by the proportion of correctly completed sets among sets attempted, with set targets for each metric. These two performance dimensions have inherent trade-offs as improving ~~in~~ accuracy necessitates spending more time on each slider to position it correctly and thus compromising on speed. The treatments will vary the difficulty of achieving the targets, which augment the probabilistic expectations of simultaneous target achievement, and the extent of explanation provided regarding the relationship between the two performance dimensions and their targets, which affect the cognitive complexity of reconciling them.

Methodology

Design

The experiment is divided into two parts: a real effort task and then a questionnaire. The former provides the main data on task performance and effort exertion to answer the research questions, whereas the latter provides covariate data for heterogeneity and robustness analysis. There are aspects of the actual experiment being conducted which depart from the ideal design due to resource and time constraints, which I address in the design outline below, including a discussion of the trade-offs, implications, and remedies.

The real effort task is a slider task which consists of a series of slider sets, and each set contains three sliders which can be moved along a scale of 0 to 100 (Figure 1). Subjects are given five minutes to work on the task. To complete a set, subjects have to drag all sliders to or past "50" on the scale. This ensures that subjects have to actually move the sliders a considerable distance in order to complete a set, hence inducing effort in the speed dimension. To *correctly* complete a set, subjects need to correctly drag every slider at its designated number (which is always equal to or greater than 50), otherwise it will be counted as mistake. This induces effort in the accuracy dimension separate from and in tension with the speed dimension, as the latter entails just dragging the sliders as fast as possible, whereas the former requires slowing down to precisely position each slider. Each set is displayed on a separate page, so having multiple sliders in each set increases the proportion of time actually spent working on the slider task by reducing the time spent on page transitions, but too many would have reduced the sensitivity of the effort measure, tasks (correctly) completed, to effort exerted, so I decided on three. On every page during the task, subjects are shown key task performance metrics, including their time spent working up till the current set, number of tasks completed per minute, total tasks completed, and total actual mistakes made. Scoring is at the set level instead of the slider level so that the metrics have smaller quantities and can be more easily processed by subjects while completing the task.



*Please drag the sliders to their designated numbers for each row.

Figure 1: Example of a set in the slider task

The slider task was selected because it is mundane and repetitive, hence reasonably incurring a positive effort cost. This combined with the fact that working on the task provides no intrinsic value should make it inert to variation in personal motivation regarding the task. The task is also easy and intuitive so task performance will be less affected by differences in intelligence and education/ training among subjects, and more clearly maps from effort exertion. These will help to reduce noisiness in the effort measures. The task is also intentionally abstract in nature and generic in its assessment since this study seeks to find universal decision-making processes regarding effort exertion which has some generalisability to a broad range of jobs and possibly beyond the labour supply domain. The short task duration and low stakes may be unrealistic to long-term jobs, but have to be imposed for practical reasons, and still offers insight into how people respond to such multi-dimensional targets at the task level of a job (e.g. a single ride for an Uber driver), which can plausibly be aggregated to the job level. Overall, the objective is not to exactly capture how people work under targets in their specific jobs, but identify fundamental decision-making mechanisms which can apply across the various domains but also vary across them through the parameters in the model.

There are four treatment groups which vary the task parameters in terms of the reference points (i.e. targets) and how the work was assessed. Reference points will be set in the two performance dimensions: speed as measured by the number of tasks completed per minute, and accuracy as measured by the proportion of *recorded* mistakes out of completed tasks. Work is assessed by either of two critera: strict which records all actual mistakes made, and lenient which records only a quarter, the probability for each criteria differs across treatments. Subjects who are more likely to be assessed by a strict criteria thus have a lower likelihood of achieving both reference points concurrently. To reinforce this perception, subjects will be primed to think that "achieving both targets [is] manageable under a lenient criterion but highly challenging under a strict criterion". Reference points are also either presented as is or explained in greater detail by explicating how performance in the speed dimension relates to that in the accuracy dimension. Subjects who receive an explanation will be told the additional number of actual mistakes they could make under each criteria for every additional set completed per minute, and provided a table showing the maximum number of total actual mistakes allowed under each criteria for different number of total tasks completed. This is intended to reduce the cognitive complexity of reconciling both reference points.

Treatment 1 is the control with no reference points (and hence no explanation) and certainty of being assessed by a strict criterion. Treatments 2, 3, and 4 are treated groups which are set the same reference points: 9 tasks completed per minute and 10% recorded mistakes[2]. Treatments 2 and 4 have a 75% probability of getting a lenient assessment criteria and 25% probability of strict, whereas treatment 3 has the inverse. Treatments 2 and 3 have the reference points explained in greater detail, whereas treatment 4 does not. The control allows for verification of the existence of reference point effects, which is a prerequisite to identifying changes in those effects due to treatment. Comparing treatments 2 and 3 demonstrates the role of probabilistic expectations of concurrent achievement of targets in attenuating reference point effects, whereas comparing treatments 2 and 4 elicits the role of cognitive complexity. Table 1 summarises the four treatment groups and their treatment conditions.

Table 1: Treatment groups and conditions

Treatment	Reference Points	Assessment Criteria Probabilities	Explanation
1	No	100% strict	Not applicable
2	Yes	25% chance of strict, 75% chance of lenient	Yes
3	Yes	75% chance of strict, 25% chance of lenient	Yes
4	Yes	25% chance of strict, 75% chance of lenient	No

After the task, all subjects will be requested to complete an optional questionnaire on their

[2] The targets were calibrated from an initial trial of the task, such that achieving both targets under the strict criterion had zero probability in the empirical distrbution but had median probability under the lenient criterion. Subjects were from my social circle and requested to do as many sets and as accurately they could.

charcateristics and reflections on the task, providing data for robustness checks and heterogeneity analysis of treatment effects, as well as verification of the experiment's construct validity. Characteristics collected includes gender, race, age, household income, education level, economic study at the undergraduate level or higher, prior participation in behavioral science lab experiments, mouse usage in the task, computer usage frequency in general, industry of employment and/or occupational type, and loss aversion levels. These represent factors which may affect task performance aside from e[illegible] exertion so I can check for baseline balance and control for them if imbalance is found, as well as factors which may drive differential responses to the treatment so I can conduct heterogeneity analysis. Reflections on the task asks about subjects' goals for speed and accuracy, whether they attempted to achieve the set targets, and if not their reasons for ignoring the targets, which provides a qualitative check of how they interpreted the reference points and treatment conditions.

Samples will be drawn from two populations: undergraduate students at the *University of Chicago* recruited through the instructors of specific courses (TBD), and the general public of Chicago recruited through the research laboratories of the *Roman Family Center for Decision Research* (RFCDR) at the *University of Chicago's Booth School of Business*. I anticipate concerns about the external validity of the study due to this. The undergraduate student sample was chosen as a cost-free way to obtain a sufficiently large sample, but the findings may be less representative of decision-making processes in the general population. The general public sample seeks to resolve this, but due to uncertainty about the amount of funding available from RFCDR, it may be unable to collect enough data to identify treatment effects [3]. Furthermore, since these participants are recruited via RFCDR,

[3] Preliminary power analysis indicated an upper bound sample size requirement of 179 observations per treatment group (716 observations total) given a conservative estimate of the minimum detectable effect size (Pearson's r) of -0.24 standard deviations and equal outcome variances across groups.

there could still be selection bias as the composition of people who are exposed and responsive to the organisation could differ from those who are in the general population (e.g. wealthier, more educated, greater interest and familiarity with behavioral research etc). Moreover, the composition of people who elect into the study may differ from those who do not (e.g. more risk-averse, more leisure time, more motivated towards knowledge creation etc). However, in the same vein of reasoning above, I would argue that the experiment investigates general decision-making processes which has broad transference across populations, offering insight into fundamental decision rules which can set the groundwork for thinking about and modelling more specific contexts.

To incentivise participation, those from the undergraduate student sample will be offered x% (TBD) of course credit just for participating in the study, whereas those from the general public sample will be offered a flat fee for participation (technically payment is pro-rated based on time spent but task duration is fixed). Ideally, there would have been additional incentives (beyond intrinsic motivation from the targets) for effort exertion in the real effort task to better parallel workplace settings. However, this was not feasible in the student sample due to fairness concerns as awarding additional class credit based on task performance would depend on the assessment criteria which was assigned by chance, nor in the public sample due to sample size requirements and budgetary constraints. Again, this may cause selection issues; Harrison, Lau, and Elisabet Rutström (2009) evinced that a fixed payment for participation can attract more risk-averse people, so this will be important to keep in mind when evaluating treatment effects.

Participants will complete the experiment virtually on Qualtrics. Conducting the experiment in-person would have afforded greater control over the task environment and hence reduced noise in effort measures, but given the ~~paper's~~ time constraints, I opted for an online mode to improve accessibility so that I could more quickly collect sufficient data. Furthermore, in-person experiment conduct is particularly difficult to operationalise for students, as I need to conduct the experiment outside of class time which is difficult to organise given the different schedules of the students, which would necessitate running the experiment on several occasions, and obtaining permission to use university facilities each time. Given limited time, this would be ill-advised, and while implementation was more viable at the RFCDR labs, it was advisable to standardise the experiment medium for better comparability across the two subject pools since that was the whole point of supplementing with the student sample.

Treatment assignment is completely randomised at the individual subject level. A stratified randomisation scheme would have been preferred to mitigate against possible covariate imbalance between treatment groups which can occur chance, with an optimal matched pair design minimising the mean-squared error of treatment effect estimates conditional on covariates (Bai 2022). However, this would require collection of covariate data pre-treatment for sorting/ matching. Currently, this information is acquired through a post-task questionnaire, which is optional and unincentivised due to budgetary constraints. It is likely that imposing such a requirement pre-task without additional incentives will deter participation and have a counterproductive effect on statistical inferential power and precision. However, with sufficient finances to incentivise and time to conduct a pre-treatment survey, I would have adopted stratified randomisation by blocking on gender, race, age, household income, education level, whether the subject studied economics at the undergraduate level, prior participation in behavioral science lab experiments, frequency of computer usage, industry of employment and/or occupational type, and loss aversion level ~~(method of estimation outlined in the analysis plan)~~.

Since costs do not vary treatment groups given the incentive structure, optimal sample size allocation across treatment groups is solely dependent on outcome variances according to $\pi_1/\pi_2 = \sigma_1/\sigma_2$ where π_i is the sample size proportion of treatment group i and σ_i is the outcome standard deviation of group i. However, the original KR model and my appended model predicts different outcome variance ratios between the treatment groups (see subsection below). Both models agree that treatment 1 will have the highest effort variance due to the absence of targets and their anchoring effects. However, the models disagree on the effort variances of the other treatment groups. The original KR model predicts that treatments 2 and 4 will have the same effort variance, and treatment 3 will have a lower outcome variance due to the greater likelihood of being assessed by a strict criteria inducing greater effort to not make mistakes. Conversely, my appended model predicts that treatment 2 will have the lowest effort variance, and treatments 3 and 4 will have higher effort variance due to attenuation of the targets.[4] Normalising standard deviation in treatment group 1 to be 1, and assuming the targets reduce the standard deviation in treatment 2 by 20%[5], and bounding the strict effect to be another 20% reduction, Table 2 shows the standard deviation and hence sample size ratios stipulated by the two models, and the one chosen by compromising between them.

[4] It is also possible that treatment 1 has the lowest variance if subjects mostly choose not to exert any effort since there is no additional incentive to do so, but this seems unlikely based on results from preliminary data. While subjects were mainly from my social circle and may have felt obligated to perform better, I believe the typical person would still work on the task given the short duration and that they have to remain on the screen.

[5] This is based on estimates from my undergraduate lab experiment which compared effort exertion under a non-financial target and none. Although the task was different and effort was uni-dimensional, the setting still bears similarities to this one so the estimates have portability.

Table 2: Predicted effort variances

Treatment	KR prediction	Appended prediction	Compromise
1	1	1	1
2	0.8	0.7	0.8
3	(0.64, 0.8)	(0.8, 1)	0.8
4	0.8	(0.8, 1)	0.9

Theoretical specification and hypotheses

I develop the refinements to the KR model to capture the effect of probabilistic expectations and cognitive complexity in reference-dependence, which produce distinct testable predictions for the experiment.

In the experiment, the agent works on a task where he/she has to exert effort e, and has targets N for the number of tasks completed per minute, and Q for the percentage of *actual* mistakes. e is split into e_1, effort in speed, and e_2, effort in accuracy. First, consider a simplified version where outcomes are deterministic, reference points are degenerate, and gain-loss utilities are linear with constant loss aversion. Under the KR model, expected utility from effort across two dimensions is given by the KR model as

$$
\begin{aligned}
U = & p(e_1, e_2) - c(e_1, e_2) + \\
& \mu_1[(n(e_1) - N)\mathbb{I}(n \geq N) + \lambda_1(n(e_1) - N)\mathbb{I}(n < N)] + \\
& \mu_2[(Q - q(e_2))\mathbb{I}(q \leq Q) + \lambda_2(Q - q(e_2))\mathbb{I}(q > Q)]
\end{aligned}
\tag{1}
$$

$p(e)$ is the level payoff from effort exertion, summed across both dimensions. $c(e)$ is the cost of effort. $\mu_1[(n(e_1) - N)\mathbb{I}(n \geq N) + \lambda_1(n(e_1) - N)\mathbb{I}(n < N)]$ is the gain-loss utility in the speed dimension, where $\mu_1 \geq 0$ is the gain-loss parameter, $\lambda \geq 1$ is the loss aversion parameter, and $\mathbb{I}(.)$ is an indicator function equaling 1 when the condition in the bracket holds and 0 otherwise. $\mu_2[(Q - q(e_2))\mathbb{I}(q \leq Q) + \lambda_2(Q - q(e_2))\mathbb{I}(q > Q)])]$ is analogously defined for the accuracy dimension.

To account for the role of probabilistic expectations and cognitive complexity in reference point effects, I propose the appended model

$$
\begin{aligned}
U = & p(e_1, e_2) - c(e_1, e_2) + \\
& E[\mathbb{P}(\{n \geq N - \varepsilon\} \cap \{q \leq Q + \varepsilon\})] \times \theta \times \quad \{\mu_1[(n(e_1) - N)\mathbb{I}(n \geq N) + \lambda_1(n(e_1) - N)\mathbb{I}(n < N)] + \\
& \mu_2[(Q - q(e_2))\mathbb{I}(q \leq Q) + \lambda_2(Q - q(e_2))\mathbb{I}(q > Q)]\}
\end{aligned}
\tag{2}
$$

The first additional term $E[\mathbb{P}(\{n \geq N - \varepsilon\} \cap \{q \leq Q + \varepsilon\}]$ captures the agent's expected probability of simultaneously achieving (within some bandwidth ε of) the reference points. When this expected probability is lower, the agent weights the gain-loss utilities less and hence is less responsive to the reference points. The second additional parameter $\theta \geq 0$ is a parameter decreasing in the cognitive complexity required to integrate the multiple reference points, so greater cognitive complexity attenuates reference point effects.

Extending the model to the context of the slider task with strict and lenient assessment criteria, we have

$$\begin{aligned} U =& p(e_1, e_2) - c(e_1, e_2) + \\ & \mathbb{P}^E \times \theta \times [\phi_1(\mu_1, \lambda_1, n(e_1), N) + \phi_2(\mu_2, \lambda_2, n(e_2), Q)] \end{aligned}$$

Where

$$\begin{aligned} \phi_1 =& \mu_1[(n(e_1) - N)\mathbb{I}(n \geq N) + \lambda_1(n(e_1) - N)\mathbb{I}(n < N)] \\ \phi_2 =& \mu_2\{P_s[(Q - q(e_2))\mathbb{I}(q \leq Q) + \lambda_2(Q - q(e_2))\mathbb{I}(q > Q)] + \\ & P_l[(4Q - q(e_2))\mathbb{I}(q \leq Q) + \lambda_2(4Q - q(e_2))\mathbb{I}(q > Q)]\} \end{aligned} \tag{3}$$

$\mathbb{P}^E$ is the expected probability term from before, P_s is the probability of getting a strict assessment criteria, and $P - l$ is the probability of getting a leninet criteria. Differentiating with respect to effort, the two models provide distinct predictions for optimal effort provision in the real effort experiment[6]. Essentially, without accounting for the role of probabilistic expectations and cognitive complexity, the KR model predicts that subjects would respond to a higher chance of being assessed by a strict criteria by reducing actual mistakes made since they are more likely to be recorded, exerting more effort in

[6] Refer to appendix for formal derivation of the first-order conditions.

the accuracy dimension either in addition to effort in the speed dimension or at the expense of it, and subjects' responsiveness to the targets are not affected by whether there is an explanation of how the two performance dimensions and their targets are related. Conversely, the appended model predicts that subjects faced with a higher chance of being assessed by a strict criteria would exert less effort in both performance dimensions since they believe it less likely to achieve them and hence attenuate them, and subjects provided with an explanation would exert more effort in both performance dimensions since they are better able to reconcile the targets to inform their effort exertion choices and hence act more responsively to the targets. These produce testable hypotheses about the treatment effects, holding the assessment criteria constant.

KR model predictions:

- KR1: Treatments 2 and 4 will have similar positive effects on the probability of achieving any target.
- KR2: Treatment 3 should have a larger positive effect than treatments 2 and 4 for achieving the accuracy target and either equal or lower positive effect for achieving the speed target.

Appended model predictions:

- A1: Treatment 3 will have lower positive effect for achieving any target than treatment 2, and in the extreme tend to treatment 1 (no effect).
- A2: Treatment 4 will have a lower positive effect for achieving any targets than treatment 2, and in the extreme tend to treatment 1 (no effect).

Analysis plan

Regression of target achievement probabilities

The main analysis will be via a multinomial logistic regression, in which the saturated model to be estimated is

$$log\frac{\mathbb{P}(Y=y)}{\mathbb{P}(Y=y_0)} = \alpha + \beta D^T + \theta criterion + \eta X + \varepsilon_i$$

Y is the categorical variable for whether subjects achieved the targets, with four possible values: achieving both targets, achieving the speed target only, achieving the accuracy target only, and achieving none. Achieving none is set as the baseline (i.e. y_0). D^T is the vector of treatment group dummies, with treatment 1 as the baseline (i.e. omitted). $criterion$ is the assessment criteria, with strict as the baseline. X is the vector of covariates collected in the post-task survey. The estimated coefficients are readily interpretable as the increase in the log odds ratio of achieving both targets/ only the speed target/ only the accuracy target relative to achieving no targets, and hence map nicely to treatment effects on the probability of target achievement, which in turn measures the strength of reference point effects.

The KR model predicts that all β_i will be significant, with $\beta_2 = \beta_4 < \beta_3$ for accuracy and either all equal or $\beta_2 = \beta_4 > \beta_3$ for speed, whereas the appended model predicts $\beta_2 > \beta_3$ and $\beta_2 > \beta_4$ for any target category, and in fact β_3 and β_4 may be insignificant. In interpreting significance of results, I will need to correct for multiple hypothesis testing since I am essentially estimating three models with different dependent variables corresponding to each category of target achievement.[7] Since the

[7] While I am also testing for multiple treatment effects, I do not consider them under the same family of hypotheses, as the existence of probabilistic expectation effects should not affect that of cognitive complexity effects. In other words, they represent separate lines of inquiry which can stand on their own.

categories of target achievements are likely correlated, I will use the List, Shaikh, and Xu (2019) and List, Shaikh, and Vayalinkal (2023) method which controls the family-wise error rate while allowing for more powerful detection of departures from the null by accounting for such correlations and leveraging covariate data.

Prior to running the regression, I will check for baseline balance in the covariates between the treatment groups using individual Kruskal-Wallis rank sum tests for each covariate and a joint likelihood ratio test for the regression of treatment groups on the covariates. This will help to determine whether it is important to control for the covariates in the regression. I will also check for substitution effects between the two effort dimensions which will inform which subhypotheses I should focus on.

To check that outcome differences between treatments are indeed stemming from the reference points, I will run the additional regression

$$log\frac{\mathbb{P}(Y = y)}{\mathbb{P}(Y = y_0)} = \alpha + \beta D^T + \gamma(D_2 * L) + \theta criterion + \eta X + \varepsilon_i$$

L is loss aversion level, and if treatment effects in treatment 2 are truly coming from the targets, they should increase in the loss aversion level, so γ should be significant. Loss aversion will be measured in the manner of Campos-Mercade et al. (2024), with subjects having been asked to indicate the number of slider sets they are willing to complete under fixed and random piece rates.

Next, for robustness checks, I will redefine the categories for target achievement using different bandwidths around the target to see whether the significance of findings depend on arbitrary alterations in the effort measure. I will also compare whether similar treatment effects are found for the student sample, public sample, and pooled sample, and discuss any differences with respect to sampling bias.

The final point of interest is heterogeneity analysis. I will to examine (robust) significant differences in baseline effort exertion/ task performance across subgroups with different characteristics (i.e. significant coefficient on the corresponding regressor) and significant differences in treatment effects across subgroups (significant coefficient on the interaction term between the regressor and the treatment dummy).

Bibliography

Abeler, Johannes, Armin Falk, Lorenz Goette, and David Huffman. 2011. "Reference Points and Effort Provision." *American Economic Review* 101 (2): 470–92. https://doi.org/10.1257/aer.101.2.470.

Bai, Yuehao. 2022. "Optimality of Matched-Pair Designs in Randomized Controlled Trials." *American Economic Review* 112 (12): 3911–40. https://doi.org/10.1257/aer.20201856.

Campos-Mercade, Pol, Lorenz Goette, Thomas Graeber, Alexandre Kellogg, and Charles Sprenger. 2024. "De Gustibus and Disputes about Reference Dependence." R\&{{R}} at {{The Review}} of {{Economic Studies}}.

Crawford, Vincent P, and Juanjuan Meng. 2011. "New York City Cab Drivers' Labor Supply Revisited: Reference-Dependent Preferences with Rational-Expectations Targets for Hours and Income." *American Economic Review* 101 (5): 1912–32. https://doi.org/10.1257/aer.101.5.1912.

Genesove, D., and C. Mayer. 2001. "Loss Aversion and Seller Behavior: Evidence from the Housing Market." *The Quarterly Journal of Economics* 116 (4): 1233–60. https://doi.org/10.1162/003355301753265561.

Harrison, Glenn W., Morten I. Lau, and E. Elisabet Rutström. 2009. "Risk Attitudes, Randomization to Treatment, and Self-Selection into Experiments." *Journal of Economic Behavior & Organization* 70 (3): 498–507. https://doi.org/10.1016/j.jebo.2008.02.011.

Hayes, Robert H. 1984. *Restoring Our Competitive Edge : Competing Through Manufacturing*. Edited by Steven C. Wheelwright. New York: Wiley.

Hayes, Robert H, and Roger W Schmenner. 1978. "How Should You Organize Manufacturing." *Harvard Business Review* 56 (1): 105–18.

Heath, Chip, Richard P. Larrick, and George Wu. 1999. "Goals as Reference Points." *Cognitive Psychology* 38 (1): 79–109. https://doi.org/10.1006/cogp.1998.0708.

Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler. 1990. "Experimental Tests of the Endowment Effect and the Coase Theorem." *Journal of Political Economy* 98 (6): 1325–48. https://doi.org/10.1086/261737.

Kahneman, Daniel, and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision Under Risk." *Econometrica : Journal of the Econometric Society* 47 (2): 263. https://doi.org/10.2307/1914185.

Koszegi, B., and M. Rabin. 2006. "A Model of Reference-Dependent Preferences." *The Quarterly Journal of Economics* 121 (4): 1133–65. https://doi.org/10.1093/qje/121.4.1133.

List, John A., Azeem M. Shaikh, and Atom Vayalinkal. 2023. "Multiple Testing with Covariate Adjustment in Experimental Economics." *Journal of Applied Econometrics* 38 (6): 920–39. https://doi.org/10.1002/jae.2985.

List, John A., Azeem M. Shaikh, and Yang Xu. 2019. "Multiple Hypothesis Testing in Experimental Economics." *Experimental Economics* 22 (4): 773–93. https://doi.org/10.1007/s10683-018-09597-5.

Plott, Charles R, and Kathryn Zeiler. 2005. "The Willingness to Pay–Willingness to Accept Gap, the 'Endowment Effect,' Subject Misconceptions, and Experimental Procedures for Eliciting Valuations." *American Economic Review* 95 (3): 530–45. https://doi.org/10.1257/0002828054201387.

Skinner, Wickham. 1974. "The Focused Factory." *Harvard Business Review* 52: 113–21.

———. 1996. "MANUFACTURING STRATEGY ON THE 'S' CURVE." *Production and Operations Management* 5 (1): 3–14. https://doi.org/10.1111/j.1937-5956.1996.tb00381.x.

Swamidass, Paul M., and Neil R. Darlow. 2000. "FOCUSED FACTORY." In *Encyclopedia of Pro-*

duction and Manufacturing Management, edited by P. M. Swamidass, 219–24. New York, NY: Springer US. https://doi.org/10.1007/1-4020-0612-8_355.

Tversky, A., and D. Kahneman. 1991. "Loss Aversion in Riskless Choice: A Reference-Dependent Model." *The Quarterly Journal of Economics* 106 (4): 1039–61. https://doi.org/10.2307/2937956.

Von Rechenberg, Tobias, Dominik Gutt, and Dennis Kundisch. 2016. "Goals as Reference Points: Empirical Evidence from a Virtual Reward System." *Decision Analysis* 13 (2): 153–71. https://doi.org/10.1287/deca.2016.0331.

Appendix

Derivation of first-order conditions for experimental theoretical specification

Beginning from the utility function,

$$
\begin{aligned}
U =& p(e_1, e_2) - c(e_1, e_2) + \\
& \quad \mathbb{P}^E \times \theta \times [\phi_1(\mu_1, \lambda_1, n(e_1), N) + \phi_2(\mu_2, \lambda_2, n(e_2), Q)]
\end{aligned}
$$

Where

$$
\begin{aligned}
\phi_1 =& \mu_1[(n(e_1) - N)\mathbb{I}(n \geq N) + \lambda_1(n(e_1) - N)\mathbb{I}(n < N)] \\
\phi_2 =& \mu_2\{P_s[(Q - q(e_2))\mathbb{I}(q \leq Q) + \lambda_2(Q - q(e_2))\mathbb{I}(q > Q)] + \\
& \quad P_l[(4Q - q(e_2))\mathbb{I}(q \leq Q) + \lambda_2(4Q - q(e_2))\mathbb{I}(q > Q)]\} \qquad (4)
\end{aligned}
$$

we can split it into the level component, gain-loss component in the speed dimension, and gain-loss component in the effort dimension, and differentiate each component with respect to effort. First consider the original KR model without $\mathbb{P}^E$ and θ.

The derivative of the level component is uninteresting across treatments as it retains the same functional form as effort exertion and task performance varies.

The matrix of partial derivatives for the gain-loss component in the speed dimension are given by

$$n \geq N : [\mu_1 n'(e_1),\ \mu_1 n'(e_1)\frac{de_1}{de_2}]$$

$$n < N : [\mu_1 \lambda_1 n'(e_1),\ \mu_1 \lambda_1 n'(e_1)\frac{de_1}{de_2}]$$

Note the discontinuity in the marginal utility of exerting effort at the target N. The agent is more incentivised to exert effort in the speed dimension when below the speed target N than above due to loss aversion as captured by $\lambda > 1$. If there are substitution effects between the the two effort dimensions, i.e. $\frac{de_1}{de_2} < 0$, then the inverse will hold true for effort in the accuracy dimension relative to N.

The derivatives for the gain-loss component in the accuracy dimension are given by

$$q \leq Q : [\mu_2 |q'(e_2)| \frac{de_2}{de_1},\ \mu_2 q'(e_2)]$$

$$Q < q \leq 4Q : [\mu_2 (P_s \lambda_2 + P_l) |q'(e_2)| \frac{de_2}{de_1},\ \mu_2 (P_s \lambda_2 + P_l) q'(e_2)]$$

$$q > 4Q : [\mu_2 \lambda_2 |q'(e_2)| \frac{de_2}{de_1},\ \mu_2 \lambda_2 |q'(e_2)|]$$

Similar insights apply for the accuracy dimension, except there are two discontinuities, one corresponding to the achieving the accuracy target under the strict criterion, and the next corresponding to achieving the accuracy target under the lenient criterion. Increasing the probability of getting a strict assessment criteria thus further incentivises agents to strive for the strict accuracy target, leading to a higher probability of achieving the accuracy target holding the assessment criteria constant.

At the optimum,

$$c'(e_1, e_2) = p'(e_1, e_2) + \phi_1'(e_1, e_2) + \phi_2'(e_1, e_2)$$
$$\frac{MC(e_1)}{MC(e_2)} = \frac{MB(e_1)}{MB(e_2)}$$

Thus, the KR model predicts that when substitution effects between the two effort dimensions are negligible, subjects in treatment 3 are more likely to achieve the accuracy target and equally likely to achieve the speed target as compared to treatment 2, or if substitution effects are appreciable, then subjects in treatment 3 are more likely to achieve the accuracy target at the expense of the speed target. The KR model does not discriminate effects between treatments 2 and 3.

However, when we add the additional parameters, then the first-order conditions become

$$c'(e_1, e_2) = p'(e_1, e_2) + \mathbb{P}^E \times \theta \times [\phi_1'(e_1, e_2) + \phi_2'(e_1, e_2)]$$
$$\frac{MC(e_1)}{MC(e_2)} = \frac{MB(e_1)}{MB(e_2)}$$

Treatment 3 lowers $\mathbb{P}^E$ whereas treatment 4 lowers θ, thus attenuating the reference point effects and leading to less likely target achievement in either dimension relative to treatment 2.