

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

Aspen Han

Department of Economics, University of Chicago

**Author Note**

Aspen Han  <https://orcid.org/0000-0003-1474-7968>

Correspondence concerning this article should be addressed to Aspen Han, Department of Economics, University of Chicago, 1126 E. 59th Street, Chicago, IL 60637, Email: [xiangyuhan@uchicago.edu](mailto:xiangyuhan@uchicago.edu)

**Abstract**

This is the abstract.

*Keywords:* conflicting multi-dimensional reference points, probabilistic expectations, cognitive complexity

### **Too hard to get: the role of probabilistic expectations and cognitive complexity in destructive multi-dimensional reference points**

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 & Schmenner, 1978](#); [Skinner, 1974, 1996](#); [Swamidass & 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 et al., 1999](#); [Von Rechenberg et al., 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?

My research is theoretically founded on the Koszegi and Rabin ([2006](#)) model of reference-dependence (henceforth KR model). Reference points have redefined preference

modeling in economics. Introduced as a core component of Kahneman and Tversky's (1979; 1991) prospect theory, it posits that people evaluate outcomes relative to a reference point rather than on absolute terms and weight losses more than gains. However, they did not identify the source of reference points, which became a source of contention. The KR model endogenises 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 & Mayer, 2001; Kahneman et al., 1990) and refutes to it (e.g. Plott & Zeiler, 2005; Tversky & Kahneman, 1991). Pinpointing the source of reference points was a major contribution as it allowed for more detailed studies into their effects and design, which motivates my use of the KR model as a theoretical baseline. The KR model also partially reconciles the EU theorem with prospect theory as KR considers the 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 it weights outcomes by their objective probabilities. This enables the KR model to satisfy internal consistency axioms such as transitivity which strengthens its normative appeal. However, the KR model, similar to most if not all reference point and EU models, also assumes that utilities across different dimensions of consumption are additively separate, which I seek to challenge. Yet, it seems unrealistic to think that people would view reference points in isolation from one another and determine how much to work towards each with complete disregard for the others.

Beyond the hypothetical examples and theoretical framework, my research builds upon empirical studies which have applied the KR model. 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. 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 test 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 and 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 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 given the logistical constraints.

In the experiment, subjects worked on a real effort task where they had 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, and were set targets for each metric. These two performance dimensions had inherent trade-offs as improving in accuracy necessitated spending more time on each slider to position it correctly and thus compromising on speed. The treatments varied the difficulty of achieving the targets, which augmented the probabilistic expectations of simultaneous target achievement, and the extent of explanation about the relationship between the two performance dimensions and their targets, which affected the cognitive complexity of reconciling them.

## **Methods**

### **Design and execution**

The experiment was divided into two parts: a real effort task and then a questionnaire. The former provided the main data on effort exertion to answer the research questions, whereas the latter provided covariate data for heterogeneity and robustness analysis.

The real effort task was a slider task which consisted of a series of slider sets, and each set contained three sliders which could be moved over a scale of 0 to 100. Subjects were given five minutes to work on the task. To complete a set, subjects had to drag all sliders to or past the “50” point mark. This ensured that subjects had 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 needed to correctly position every slider at its designated number (which was always equal to or greater than 50), otherwise it could be counted as mistake. This induced effort in the accuracy dimension. Each set was displayed on a separate page, so having multiple sliders in each set increased 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 tasks completed to effort exerted and in turn the granularity of the effort measure, so I decided on three. On every

page during the task, subjects were shown key task metrics, including their time spent working on all previous sets, number of tasks completed per minute, total tasks completed, and total actual mistakes made. Measurement of effort was at the set level instead of the slider level so that task metrics had smaller quantities and could be more easily processed by subjects while completing the task.

The slider task was selected as it was mundane and repetitive, hence reasonably incurring a positive effort cost. This combined with the fact that working on the task provided no intrinsic value should have also made it inert to variation in personal motivation regarding the task. The task was easy and intuitive so performance on it would be less affected by differences in intelligence and education/ training among subjects. The task was also intentionally more abstract and the skills assessed were also generic since the experiment sought to find out general decision-making processes regarding effort exertion which could be generalizable to a broad range of jobs. Finally, the short task duration may not be realistic to how people optimise with respect to targets in different performance dimensions for a long-term job, but had to be imposed due to budgetary constraints, and it 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 could be aggregated to the job level.

Subjects were randomly assigned at the individual level into four treatment groups, which varied the slider task in terms of the reference points (i.e. targets) and how the work was assessed. Reference points were 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 in completed tasks. Work was assessed by either of two criteria: strict which recorded all actual mistakes made, and lenient which recorded only a quarter. Subjects who were more likely to be assessed by a strict criteria thus had a lower likelihood of achieving both reference points concurrently. To reinforce this perception, subjects were primed to think that “achieving both targets [was] manageable under a lenient criterion but highly challenging under a strict criterion”. Reference points were also either presented as is or explained in greater detail by mapping

performance in the speed dimension to that in the accuracy dimension: subjects who received an explanation were 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 was intended to reduce the cognitive complexity of reconciling both reference points.

Treatment 1 was the control with no reference points (and hence no explanation) and certainty of being assessed by a strict criterion. Treatments 2, 3, and 4 were set the same reference points: 9 tasks completed per minute and 10% recorded mistakes, and primed. Treatments 2 and 4 had a 75% probability of getting a lenient assessment criteria and 25% probability of strict, whereas treatment 3 had the inverse. Treatments 2 and 3 had the reference points explained in greater detail, whereas treatment 4 did not. The control allows for verification of the existence of reference point effects, which is a prerequisite to identifying any changes in those effects. Comparing treatments 2 and 3 demonstrates the role of low probabilistic expectations of achievement 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, subjects were requested to complete an optional questionnaire on their



demographics, reflections on the task, and loss aversion, providing data for heterogeneity analysis and robustness checks of treatment effects and verification of the experiment's construct validity. Demographic information collected included gender, race, age, household income, educational level, and whether subjects studied economics at the undergraduate level or above. Reflections on the task asked about subjects' goals for speed and accuracy during the task, whether they attempted to achieve the set targets, and if not their reasons for ignoring the targets, which provided a qualitative check of whether the reference points had inherent effects and the potential treatment effects. This section also asked whether they used a mouse for the task, which could have caused differences in task performance not attributed to effort exertion. Finally, subjects indicated the number of slider sets they were willing to complete under fixed and random piece rates, and by comparing the differences in the indicated number between fixed and random piece rates with the same expected payment, I can estimate subjects' loss aversion levels. This allowed for loss aversion measures specific to the slider task and more broadly the real effort/ labor supply domain, although it was in terms of pecuniary incentives instead of non-pecuniary targets but it was difficult to elicit the latter through self-reports, and there could still be variation between different tasks within the real effort domain.

Samples were 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*. The former was chosen to alleviate financial costs and the latter was chosen for better representativeness of decision-making in the general population.

Participants completed 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<sup>1</sup>. Furthermore, in-person

---

<sup>1</sup> Preliminary power analysis had indicated a sample size requirement of 134 observations per treatment group (536

experiment conduct was also particularly difficult to operationalise for students given the limited time, as I needed to conduct the experiment outside of class time which was 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. While implementation was more viable at the RFCDR labs, it was advisable to conduct the experiment online for both for better comparability across the two subject pools.

To incentivise participation, Students were offered  $x\%$  (TBD) of course credit for participating in the study, whereas Individuals were offered a flat fee for participation. Ideally, there would have been additional incentives (aside from intrinsic motivation from the targets) for effort exertion in the real effort task. 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.

### Theoretical specification and hypotheses

In the experiment, the agent works on a task where he/she has to exert effort  $e$ , and has reference points  $N$  for the number of tasks completed per minute, and  $Q$  for the percentage of mistakes made.  $e$  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

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

---

observations total) given a conservative estimate of the minimum detectable effect size (Pearson's  $r$ ) of -0.24 and equal outcome variances across groups.

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

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

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  $\varepsilon$  of) all 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 two models to the context of the slider task with strict and lenient assessment criteria, the two models provide distinct predictions for optimal effort provision in the real effort experiment<sup>2</sup>. 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 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

---

<sup>2</sup> Refer to appendix for formal derivation of the first-order conditions

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.

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 both targets or for achieving  $Q$  at the expense of  $N$ .

Appended model predictions:

- A1: Treatment 3 will have lower positive effect for achieving any target than treatment 2.
- A2: Treatment 4 will have a lower positive effect for achieving any targets than treatment 2.

## Results

### Data overview

In total, the experiment garnered 21 observations.

### Effort exertion and target achievement

### Relationship with covariates

## Discussion

### Robustness checks

### Limitations

The arguably biggest criticism of this study would be its sampling. The undergraduate student sample was selected as a convenient and cost-free way to pilot the study and potentially increase sample size, though the findings specific to this sample would be less representative of

decision-making processes with respect to multiple targets and effort exertion in the general population, especially since the students study Economics. Hence, I also sampled from the general public to alleviate this concern, although there could still be potential bias as participants were recruited and hence the composition of people who are exposed and responsive to the organisation could differ from those who are not, and of course A fixed amount of compensation was provided to incentivise participation.

**Conclusion**

## References

- Abeler, J., Falk, A., Goette, L., & Huffman, D. (2011). Reference Points and Effort Provision. *American Economic Review*, 101(2), 470–492. <https://doi.org/10.1257/aer.101.2.470>
- Crawford, V. P., & Meng, J. (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), 1912–1932. <https://doi.org/10.1257/aer.101.5.1912>
- Genesove, D., & Mayer, C. (2001). Loss Aversion and Seller Behavior: Evidence from the Housing Market. *The Quarterly Journal of Economics*, 116(4), 1233–1260. <https://doi.org/10.1162/003355301753265561>
- Hayes, R. H. (1984). *Restoring our competitive edge : Competing through manufacturing* (S. C. Wheelwright, Ed.). Wiley.
- Hayes, R. H., & Schmenner, R. W. (1978). How should you organize manufacturing. *Harvard Business Review*, 56(1), 105–118.
- Heath, C., Larrick, R. P., & Wu, G. (1999). Goals as Reference Points. *Cognitive Psychology*, 38(1), 79–109. <https://doi.org/10.1006/cogp.1998.0708>
- Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1990). Experimental Tests of the Endowment Effect and the Coase Theorem. *Journal of Political Economy*, 98(6), 1325–1348. <https://doi.org/10.1086/261737>
- Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2), 263. <https://doi.org/10.2307/1914185>
- Koszegi, B., & Rabin, M. (2006). A Model of Reference-Dependent Preferences. *The Quarterly Journal of Economics*, 121(4), 1133–1165. <https://doi.org/10.1093/qje/121.4.1133>
- Plott, C. R., & Zeiler, K. (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), 530–545. <https://doi.org/10.1257/0002828054201387>
- Skinner, W. (1974). The focused factory. *Harvard Business Review*, 52, 113–121.

- Skinner, W. (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, P. M., & Darlow, N. R. (2000). FOCUSED FACTORY. In P. M. Swamidass (Ed.), *Encyclopedia of Production and Manufacturing Management* (pp. 219–224). Springer US. [https://doi.org/10.1007/1-4020-0612-8\\_355](https://doi.org/10.1007/1-4020-0612-8_355)
- Tversky, A., & Kahneman, D. (1991). Loss Aversion in Riskless Choice: A Reference-Dependent Model. *The Quarterly Journal of Economics*, 106(4), 1039–1061. <https://doi.org/10.2307/2937956>
- Von Rechenberg, T., Gutt, D., & Kundisch, D. (2016). Goals as Reference Points: Empirical Evidence from a Virtual Reward System. *Decision Analysis*, 13(2), 153–171. <https://doi.org/10.1287/deca.2016.0331>

## **Appendix**

### **Derivation of first-order conditions for experimental theoretical specification**