NBER WORKING PAPER SERIES

ARE REFERENCE POINTS MERELY LAGGED BELIEFS OVER PROBABILITIES?

Ori Heffetz

Working Paper 24721
http://www.nber.org/papers/w24721

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2018

I thank Ned Augenblick, Keith Ericson, David Gill, Botond Kőszegi, Victoria Prowse, Charlie Sprenger, and Florian Zimmermann for useful discussions, and especially Ted O'Donoghue and Matthew Rabin for many insightful conversations; Charlie and Florian also for generously sharing their data and programs; seminar participants at Cornell, Harvard, Hebrew U, Norwegian School of Economics, Stanford, SUNY Buffalo, Technion, Tel Aviv U, UC Berkeley, and UCSD Rady for helpful comments; the RatioLab team and especially Deborah Marciano-Romm, and the Business Simulation Lab team and especially Margaret Shackell, for help running the experiments; Maya Catabi, Bnaya Dreyfuss, and Guy Ishai for research assistance, and especially Aharon Haver for thoughtful and creative research assistance throughout the entire development of the project. I also thank Danny Kahneman for triggering off the train of thought that initiated this research, when in December 2010 he wrote in an email: "These are very clean instructions but they are almost ridiculously complicated. The students might be able to figure out how to answer the questions at the end, but I don't think any psychologist would agree that this is a good manipulation of expectations." This research project was supported by the I-CORE program of the Planning and Budgeting Committee and the Israel Science Foundation (grant no. 1821/12), and by the S.C. Johnson Graduate School of Management. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Are Reference Points Merely Lagged Beliefs Over Probabilities?
Ori Heffetz
NBER Working Paper No. 24721
June 2018
JEL No. D12,D84,D91,J22

ABSTRACT

What explains the mixed evidence from laboratory tests of Kőszegi and Rabin's (2006 and later) model of expectations-based reference-dependent preferences? We investigate one hypothesis: to become (behavior-affecting) reference points, probability beliefs have to sink in—being merely lagged, as the theory requires, is not sufficient. Past experiments with conflicting findings exogenously endowed subjects with beliefs that were equally lagged, but possibly unequally sunk-in. In four experiments, whose designs replicate past KR-nonsupporting experiments, we add new sink-in manipulations that endow individuals with additional, visual/physical probability impressions. Our findings are more KR-supporting in an endowment-effect setting but not in an effort-provision setting.

Ori Heffetz
S.C. Johnson Graduate School of Management
Cornell University
324 Sage Hall
Ithaca, NY 14853
and The Hebrew University of Jerusalem
and also NBER
oh33@cornell.edu

An online appendix is available at http://nber.org/~heffetz.

1 Introduction

Kőszegi and Rabin's (2006, 2007, 2009, henceforth KR) theory of reference-dependent preferences has emerged in the past decade as a fruitful, generally applicable framework.[1] It combines the idea of loss aversion from Kahneman and Tversky's (1979) prospect theory with an explicit, testable theory of expectations-based reference, relative to which gains and losses are assessed. The model appears consistent with a diverse set of empirical observations, and has inspired an increasing number of theoretical applications. However, direct lab experimental tests of key predictions of the theory have yielded mixed results. This paper aims to investigate why.

We focus on an important subset of the relevant evidence to date. Specifically, of the observed phenomena that the theory holds potential to explain, two receive special attention in KR's first (2006) paper: the endowment effect and negative wage-elasticity of labor supply (or effort provision).[2] Perhaps naturally, the first lab experimental tests of the theory focused on these two phenomena—and so do we.

We investigate the endowment effect in section 2, and effort provision in section 3. As a preparatory step, after reviewing the basic KR (2006) framework (in section 2), in each of the two sections we first re-solve the model for one central experimental paradigm (i.e., a specific experimental setup used in past work), corresponding with one of the two phenomena. We then review lab evidence from these paradigms, arguing that it is rather mixed—perhaps more mixed than has been generally appreciated. In the course of doing so, we revisit several potential explanations, including the possibility that the theory's predictions in these experimental setups are not as robust (e.g., to the choice of equilibrium concept) as is commonly believed, and the related possibility that some of these past experiments are not as clean tests of the theory as is commonly assumed.

We argue that the picture that emerges from these preparatory steps varies across the

[1] For an excellent recent survey of the literature on reference-dependent preferences, see O'Donoghue and Sprenger (2018).

[2] KR's focus on these two phenomena is highlighted in the abstract: "Applying the model to consumer behavior, we show that willingness to pay for a good is increasing in the expected probability of purchase and in the expected prices conditional on purchase. In within-day labor-supply decisions, a worker is less likely to continue work if income earned thus far is unexpectedly high, but more likely to show up as well as continue work if expected income is high."

two paradigms. In what has come to be called the *exchange paradigm* of endowment-effect experiments, two theoretically equivalent lab setups—Ericson and Fuster (2011, henceforth EF) and Heffetz and List (2014, henceforth HL) yielded conflicting results: the former appears supportive of the model, while the latter does not. In what we term the *table-counting* paradigm of effort-provision experiments, in contrast, clean and robust supportive evidence has failed to materialize to date. The original supportive evidence of Abeler, Falk, Goette, and Huffman (2011, henceforth AFGH)—the creators of this paradigm—does not seem as clean as it appeared at first; was subsequently only weakly replicated by Camerer et al. (2016) with a substantially larger sample; and was recently found by Gneezy, Goette, Sprenger and Zimmermann (2017, henceforth GGSZ) to not easily generalize. We describe these experiments, analyze the model's predictions in each case, and argue that of all the experimental treatments that have been conducted in this paradigm, a subset of GGSZ's treatments appear the cleanest implementation-wise.

The present paper's main contribution is to offer, and experimentally investigate, one specific, previously understudied, hypothesis: **reference points are not merely *lagged* expectations, as the theory postulates; rather, they are *sunk-in* expectations**. We explain these terms shortly. In a nutshell, our hypothesis shifts focus away from a merely time-based definition of reference points towards a more perceptions-based one. Importantly, under this hypothesis, past experiments could generate different findings in spite of successfully manipulating equally lagged expectations because—and to the extent that—expectations did not sink in with subjects equally across experiments. In turn, this could have been due to, for example, differences in experiment-specific implementation details of the expectations manipulation (e.g., the use of language, visuals, repetition, etc.). We experimentally investigate this hypothesis in four new experiments—two per paradigm—by adding to past experiments new treatments that are designed to directly manipulate sink-in.

In section 2's main part, we add new sink-in treatments to the endowment-effect experimental setup of HL. We show, with two new experiments (total $N = 480$), that our hypothesis holds promise to resolve much of the apparent inconsistency across EF's and HL's findings. Specifically, under the new sink-in manipulation, behavior changes in a direction consistent with the model's prediction, though some nuances remain regarding the choice of equilibrium

concept.

In section 3's main part, we investigate the new sink-in hypothesis in the effort-provision context. Having argued in that section's preparatory part that clean and robust KR-consistent evidence in the AFGH table-counting paradigm is yet to materialize, and having found, in section 2's endowment-effect setup, that a new sink-in treatment can move results to be more KR-consistent, we import the new treatment into this paradigm. We first import a version of the new treatment that is as close as possible to the one that we used in the endowment-effect setup. Our new experiment (N = 127), modeled after what we argue is GGSZ's cleanest variant, fails to yield evidence that is more consistent with KR than past evidence in this setup, although, again, interpretations depend on the choice of equilibrium concept. We then try what we think is a stronger sink-in treatment, tailored more closely to the specific details of the setting, in our fourth and final experiment (N = 120). We again fail to find that results significantly move in the KR direction. We discuss potential explanations for the failure of this paradigm to yield clean, robust KR-consistent findings, and its contrast with the endowment-effect exchange paradigm, in our concluding discussion in section 4.

Two conclusions cautiously emerge from our investigation. First, our sink-in hypothesis seems promising. Our endowment-effect findings suggest that the designers of future experimental tests of the theory should pay careful attention to whether they manipulate sunk-in expectations or merely lagged expectations. Second, the table-counting paradigm may be problematic as a test of the theory (in concrete ways that we discuss in sections 3 and 4). Together, these two conclusions should help steer future experimental effort in what we believe are productive directions.

We close this introduction by elaborating on our main hypothesis: that reference points are *sunk in* (rather than merely *lagged*) beliefs. In KR's theory, reference points are lagged expectations, defined as "probabilistic beliefs ... held in the recent past about outcomes. ... Specifically, a person's reference point is her probabilistic beliefs about the relevant consumption outcome held between the time she first focused on the decision determining the outcome and shortly before consumption occurs."[3] KR further clarify in a footnote:

[3]These quotes are from KR (2006), pp. 1134 and 1141. We find KR's term "probabilistic beliefs" unnecessarily ambiguous: it could mean, for example, beliefs that occur with different probabilities. We prefer (and use) the term "probability beliefs," to more clearly convey that these are beliefs about probabilities

> Our theory posits that preferences depend on *lagged* expectations, rather than expectations contemporaneous with the time of consumption. This does not assume that beliefs are slow to adjust to new information or that people are unaware of the choices that they have just made—but that preferences do not instantaneously change when beliefs do. When somebody finds out five minutes ahead of time that she will for sure not receive a long-expected $100, she would presumably immediately adjust her expectations to the new situation, but she will still five minutes later assess not getting the money as a loss.

In this example, where long-held expectations are suddenly updated, a five-minute lag is not enough for moving the reference point. But in experimental tests of the theory, a five-minute lag is often all that subjects are given for moving their reference point. To the extent that some such experiments find KR-consistent evidence due to having successfully caused subjects to move their reference point, five minutes must be enough in those settings.

Under our hypothesis, what moves the reference point is not the passage of time per se, but some sense of internalization of, or getting used to, the new expectations—which we refer to as sink-in. It is not inconceivable that sink-in takes as little as a few minutes in a low-stakes, no-prior-expectations lab experiment, but takes much longer in a higher-stakes, long-held-expectations setting.

More generally, in the naturally occurring, real-world situations that the example illustrates and, more importantly, that the theory sets out to explain in the first place, time appears to be strongly *correlated* with sink-in. Indeed, it is difficult to find counter examples where time is not a great healer. But a time-based definition of reference points may cease to be useful in lab experiments, where this naturally occurring correlation is broken down by the very experimental design that aims at testing the theory—a design based on instant endowment of exogenous expectations (rather than on natural emergence of expectations over time, with experience and learning), followed by an assessment, within a few minutes, of their effect on behavior.

The new sink-in manipulations that we introduce in this paper are meant to strengthen sink-in in exactly these lab situations where sink-in through the mere passage of time is not an

(i.e., a belief is simply a probability distribution over outcomes).

option. To do so, these manipulations add visual/physical probability impressions through active engagement of subjects with repeated demonstrations of realizations of the relevant probabilities. To the extent that they are successful, our manipulations may be helpful in the context of other experiments. Importantly, we conduct our experiments using, to the extent possible, existing designs of published work. We do this for reasons that are both practical—we can pool our newly collected data with previously collected data to increase sample size where appropriate—and substantial—we believe that puzzling past results are best investigated by exploring experimental variations on the original designs that change one element at a time. As we mention in section 4, we hope that future work will push these investigations forward into new paradigms and domains that we have not yet studied.

2 The Endowment Effect (Exchange Paradigm)

2.1 The Basic KR Framework

A consumer's utility,

$$u(\boldsymbol{c}|\boldsymbol{r}) = \sum_k m_k(c_k) + \sum_k \mu(m_k(c_k) - m_k(r_k)),$$

is a function of two K-dimensional bundles, denoted by $\boldsymbol{c}$ (consumption) and $\boldsymbol{r}$ (reference).[4] Utility is separable across dimensions and consists of two components. The first, "consumption utility," corresponds to standard (reference-independent) utility. The second, "gain-loss utility," corresponds to prospect theory's reference-dependent utility, with $\mu(x) = \eta x$ for $x > 0$, and $\mu(x) = \eta\lambda x$ for $x \leq 0$. The parameter $\eta > 0$ is the weight an individual attaches to gain-loss utility, and $\lambda > 1$ is her "coefficient of loss-aversion," quantifying by how much the pain from a loss (relative to the reference point) is greater than the pleasure from a gain of equal size. Both $\boldsymbol{c}$ and $\boldsymbol{r}$ can be stochastic, and individuals maximize expected utility.

The reference $\boldsymbol{r}$ is determined endogenously, in what KR term a *personal equilibrium*

[4]For a detailed exposition and discussion, see Kőszegi and Rabin (2006). Kőszegi and Rabin (2009) generalize this basic framework and embed it within a multiple-period dynamic framework. We return to this point in section 4, and discuss how the updated framework may affect our analysis.

(PE).[5] It is a rational-expectations equilibrium in the following sense. Given a consumer's expectations regarding the state of the world—represented by a probability distribution over *choice sets*—she forms expectations regarding choice outcomes—a probability distribution over *consumption bundles*—by making a choice plan under each possible state. These expectations over outcomes are rational in that they are consistent: a consumer who holds them as her reference will indeed find that following through, by making the ex ante expected (= planned) choices, maximizes her utility.

As we will see in the next subsection, PE does not always make a sharp prediction, as there may be multiple PE: more than one plan may be optimal to follow through once the expectations it induces became the reference point. Therefore, KR also introduce a refinement: "Insofar as a person is free to make any plan so long as she will follow it through ... she will choose her favorite plan." This refinement, termed a *preferred personal equilibrium* (PPE), is indeed just a preferred PE: when more than one PE exists, a PPE is the one that maximizes ex ante expected utility. In other words, when an individual can form more than one set of expectations regarding outcomes which, once her reference, is consistent with optimal choices ex post—she will choose as her reference point the ex ante preferred one, assuming she can.

2.2 Theoretical Predictions

To apply KR's framework to an endowment-effect setup, consider a 2-dimensional consumption bundle $\boldsymbol{c} = (c_1, c_2)$, where c_1 and c_2 are consumer goods, e.g., mugs and pens. In the beginning of an experimental session, a subject is endowed with a unit of one of the items. The subject is immediately informed that with (positive) probability q she will be permitted, in the end of the experiment, to exchange her endowed item for the other item. Specifically, in the end of the experiment she will choose "keep" or "exchange," and her choice will be implemented with probability q; with probability $1 - q$ she will keep her initially endowed item regardless of her choice.

Assume w.l.g. that $m_1(0) = m_2(0) = 0$ and that the subject is initially endowed with the bundle $(1, 0)$. She knows that in the end of the experiment she will either choose "keep,"

[5] For a formal definition, see section F.1 in the appendix.

in which case she will keep her $(1,0)$ bundle with certainty, or choose "exchange," in which case she will replace her certain bundle with the lottery $\{(1,0),1-q;(0,1),q\}$.

We now analyze the subject's maximization problem. Consider first a subject who, in the beginning of the experiment, plans to choose "keep" in the end of the experiment. Accordingly, she forms expectations to consume the bundle $(1,0)$. If in the end of the experiment she indeed chooses "keep," then her reference bundle $(1,0)$ will coincide with her consumption bundle, and her utility—expected as well as realized—will just be $m_1(1)$, with no gain-loss terms. On the other hand, if in the end of the experiment she deviates by choosing "exchange," then her utility will be $m_1(1)$ with probability $1-q$ and $m_2(1)+\eta m_2(1)-\eta\lambda m_1(1)$ with probability q. It is thus straightforward to show that given expectations to keep item 1, she will indeed choose "keep" as long as

$$m \equiv \frac{m_1(1)}{m_2(1)} \geq \frac{1+\eta}{1+\eta\lambda}. \tag{1}$$

In other words, the choice "keep" is a PE as long as (1) holds.

Alternatively, consider a subject who, in the beginning of the experiment, plans to choose "exchange" in the end of the experiment. Accordingly, she forms expectations to consume the bundle $(1,0)$ with probability $1-q$ and the bundle $(0,1)$ with probability q. Given such reference bundles, it can be shown that she will indeed choose "exchange" in the end of the experiment as long as

$$m \leq \frac{1+(1-q+q\lambda)\eta}{1+((1-q)\lambda+q)\eta}. \tag{2}$$

Thus, when (2) holds, the choice "exchange" is a PE.

While (1) is independent of q, (2) is not. When q approaches 0, the RHS of (1) and (2) coincide, and a unique PE exists: choose "keep" iff (1) (and "exchange" otherwise).[6] This result illustrates KR's explanation of past endowment-effect findings, from lab setups where subjects were endowed with item 1, and did not expect to be able to exchange it later (so $q \approx 0$ seems a reasonable interpretation of the setting). If $\eta = 0$ or $\lambda = 1$, the model reduces to the standard model, which predicts that subjects will choose "keep" iff $m \geq 1$; that is, they

[6]Notice that (a) when $q = 0$ there is no choice to be made, and (b) the PE is unique except for the razor's-edge case $m = 1$, where both "keep" and "exchange" are a PE (and a PPE). In what follows, we assume away such cases.

will choose the item with higher consumption utility. But with $\eta > 0$ and $\lambda > 1$, subjects will choose "keep" over a wider range of the ratio m, that includes values below 1. For example, with $\eta = 1$ and $\lambda = 3$, subjects will choose to keep item 1 as long as $m > \frac{1}{2}$. The endowment effect will then be generated by subjects with $m \in [\frac{1}{2}, 1]$; outside this interval, subjects choose according to consumption utility.

The model can thus explain past endowment-effect findings. But so can alternative loss-aversion models that do not assign an explicit role to expectations; for example, models that posit that the reference bundle is simply the endowed bundle $(1, 0)$. To directly test the expectations channel, KR-inspired experiments varied q across conditions. What does the model predict for low versus high q?

We start with PE. As q grows from 0, a region grows for m above $\frac{1+\eta}{1+\eta\lambda}$ where both "keep" and "exchange" are a PE, and when $q = 1$ this region stretches all the way to $\frac{1+\eta\lambda}{1+\eta}$. Thus, under PE, the model predicts the $q \approx 0$ endowment effect to weakly decrease with q, and potentially even reverse. Whether and how much it actually decreases (and even reverses) depend on how subjects choose among two possible PE—a choice that increasingly many subjects face as q grows.

If subjects choose a PPE (by comparing the expected utility under either PE), then it can be shown that they will choose the unique PPE "keep" iff (1) for $q \in (0, \frac{1+\eta(\lambda+1)}{2+\eta(\lambda+1)})$, and "keep" iff $m \geq \frac{1-(1-q)\eta(\lambda-1)}{1+(1-q)\eta(\lambda-1)}$ for $q \in [\frac{1+\eta(\lambda+1)}{2+\eta(\lambda+1)}, 1]$. Thus, as q increases, the endowment effect is predicted to first remain constant at its $q \approx 0$ level until $q = \frac{1+\eta(\lambda+1)}{2+\eta(\lambda+1)}$, and then continuously and monotonically weaken with q. Notice that when q reaches 1, the PPE simplifies to "keep" iff $m \geq 1$. That is, the KR model's PPE prediction approaches the standard model's no-endowment-effect prediction as subjects approach a condition where they know in the beginning of the experiment that in the end of the experiment they will be asked to choose between the certain bundles $(1, 0)$ and $(0, 1)$.

In summary, the model predicts, first, an endowment effect at low q's regardless of which equilibrium concept, PE or PPE, applies.[7] Second, it predicts that at high q's, the effect

[7]The analysis above shows that while in PPE, the *size* of the predicted endowment effect at low q is independent of q, in PE it does in general depend on q (and on how subjects choose a PE). But as long as q remains relatively low, the model predicts, under reasonable assumptions, a similar endowment effect under PE and PPE. To bound the difference, consider the extreme case where all subjects choose "exchange" whenever it is a PE—so the PE-predicted endowment effect shrinks with q as steeply as possible. Recall that

should not be larger than the low-q effect; it could remain roughly constant, shrink, or reverse—depending on how subjects choose among multiple PE. Third, if subjects choose a PPE, then the effect would never reverse: it would remain constant as long as q is below the cutoff $\frac{1+\eta(\lambda+1)}{2+\eta(\lambda+1)}$, then start shrinking with q, and eventually disappear at $q = 1$. Notice that this cutoff can be quite high. For example, with $\eta = 1$ and $\lambda = 3$, it is $q = \frac{5}{6}$.

2.3 Past Evidence

EF (Ericson and Fuster, 2012) conducted an experiment that fits the above setting, using two slightly different designs. In both designs, they endowed subjects with a mug, and let a coin-flip determine the probability—low ($q = \frac{1}{10}$) or high ($q = \frac{9}{10}$)—with which subjects would later be permitted to trade the mug for a pen. In EF's preferred, later design, the coin was visibly flipped in front of $N = 45$ subjects, making it clear to subjects that their assignment into low or high treatment was randomly determined. 77% and 43% of mug owners, respectively, chose "keep" in the low and high treatments—a large low-high difference in a direction consistent with subjects choosing a PPE.

In contrast, in EF's earlier design—which was otherwise similar to the later design—the random assignment of $N = 63$ subjects into the low and high treatments was not transparent to the subjects: "subjects were simply told that they have a 10% or 90% probability of permitted trade, without being informed that this probability was randomly assigned." With this earlier design, EF found the opposite result: 38% and 71%, respectively, chose "keep" in the low and high treatments. As EF observe, the earlier design is confounded along the lines suggested by Plott and Zeiler (2007): if subjects do not observe that the probability with which trade is permitted—low or high—was assigned to them at random, they may mistakenly perceive their assigned probability as informative regarding the mug's value. In particular, a subject who is given an item and is allowed to trade it with only low probability, may mistakenly infer that the endowed item is of lower value than the alternative. But the effect reversal EF find is so dramatic that it suggests to them that the PPE-consistent

with $\eta = 1$ and $\lambda = 3$, the PE-predicted effect at $q \approx 0$ is generated by subjects with $m \in [\frac{1}{2}, 1]$. In comparison, even in this boundary case, at $q = \frac{1}{6}$ for example, the effect would still be generated by all subjects with $m \in [\frac{7}{11}, 1]$. In what follows, we assume that the implied difference in the predicted effect is negligible.

effect found in the cleaner design is rather fragile, and may be easily washed out by subtle implementation details. We return to this important point in section 3, when discussing evidence from effort-provision experiments that are potentially confounded similarly to EF's early design.

Contemporaneously with EF's experiments, and independently of them, HL (Heffetz and List, 2014) also conducted an exchange experiment testing KR. As we explain in detail shortly, from the point of view of KR's model, HL's experiment was theoretically equivalent to the cleaner design of EF. HL subsequently ran two additional experiments, explicitly designed to also eliminate some implementation differences (that the theory does not specify) between their original experiment and EF's, eventually reaching a substantially larger sample ($N = 560$) than EF's. (We reproduce the relevant results from these two experiments in table 1, subsection 2.5 below.) However, in spite of the clean design, the theoretical equivalence, and the elimination of some implementation differences, HL found, for the most part, a statistically insignificant but persistent *negative* low-high difference. This finding's direction is opposite to a PPE prediction, and allows HL to reject a large PPE-consistent difference. In particular, HL's 95% confidence intervals—from each of their experiments separately, and certainly when they are pooled together—always reject a 34-percentage-point difference like that found in EF's clean design. Of course, since HL cannot reject the null of no difference, their finding is consistent with subjects choosing the same PE at all q values.[8]

Of HL's three experiments, the most relevant for our purposes is a condition called "Experiment 2, More Endowment." While it qualitatively replicates the findings from HL's Experiments 1 and 3, it is the simplest setup that was purposefully designed to be comparable to EF's setup. Briefly, it is based on a 2×2 experimental design, with two "assignment" conditions (Coin-Mug vs. Coin-Pen) and two "expectations" conditions (Weak Expectations

[8]These endowment-effect experiments follow the simple item-versus-item design of Knetsch (1989)—i.e., the *exchange* paradigm. Other endowment-effect experiments follow the more complex item-versus-money design of Kahneman, Knetsch, and Thaler (1990)—i.e., the *valuation* paradigm—where willingness to accept and/or willingness to pay are elicited for an endowed item. The latter include the pioneering study by Smith (2008); Ericson and Fuster's (2011) willingness-to-accept experiment ($N = 112$) with findings that support KR's prediction; an attempt to replicate them with a larger sample ($N = 262$) by Camerer et al. (2016), with weaker KR-supportive findings (a statistically weaker ($p = 0.055$) effect with relative size 69% of the original); and the recent experiments by Goette, Harms, and Sprenger (2017), who report finding tentative support for the model, albeit with results that "are sensitive to small changes in experimental design."

vs. Strong Expectations). Participants are presented with two goods, a mug and a pen, and are asked to toss a coin which, as they subsequently learn, determines which of the two goods is assigned to them. Participants are then told that they own the assigned item: it "belongs to you as a gift to take home." However, they are also told that with probability q they will be able to exchange their item for the other item if they want to; $q = \frac{9}{10}$ and $\frac{1}{10}$, respectively, in the Weak and Strong Expectations conditions.[9] Participants are quizzed to verify that they understand the procedure and the related probabilities—hence providing evidence that correct probability beliefs have been established by participants. After the quiz, participants fill out a survey whose sole purpose is to pass time, thereby providing a (short) lag and turning said probability beliefs into *lagged* beliefs. Having completed the survey, participants choose "keep" or "exchange"; are asked exit questions; have the 10%/90% uncertainty resolved; and receive their gift accordingly. (For differences between this and HL's other experimental conditions, see appendix E.)

HL find that assignment affects choice: participants are more likely to keep the item assigned to them by the coin-flip than to trade it. But opposite to the PPE prediction and to EF's clean finding, the effect is *smaller* in the Strong than in the Weak Expectations treatment, although the difference is not statistically different from zero. In HL's Experiment 2's More Endowment condition, that opposite difference is 20 percentage points. Thus, HL find no evidence of a PPE-consistent effect of expectations on choice *when expectations are implemented as merely lagged verified probability beliefs.*

Since the contrasting findings in EF's and HL's experiments cannot be explained within the KR framework, one has to search for explanations outside the model. Such a search lacks, by definition, the advantage of being guided by the theory and, importantly, the discipline that such guidance imposes (through limiting the degrees of experimental freedom). On the other hand, it is the only way we could think of for resolving the above empirical puzzle—of contrasting findings in two clean, theory-equivalent settings. To avoid selective reporting of evidence, we report the full details of all the new treatments we tried.

[9]Recall that while EF's experiment has two expectations treatments ($q = \frac{9}{10}$ and $\frac{1}{10}$), it has no assignment treatments (all subjects are endowed with a mug). Thus, it is *after* the transparent random mug/pen assignment in HL's experiments—which serves a purpose similar to that of EF's transparent random coin flip—that HL's design is theoretically equivalent to EF's clean design, and that the two fit the theoretical setup analyzed in subsection 2.2 above.

Our new treatments are designed to explore the hitherto untested hypothesis that lagged beliefs may not in themselves be sufficient for establishing a reference point. Rather, the lagged beliefs need to be accompanied by a subjective component, through which choice is actually affected. Simply put, said beliefs need to sink in. If this hypothesis is at least somewhat true, then there must be *something* in EF's setup whose absence in HL's setup could explain the different findings. HL present evidence that the missing component could not have been mere understanding of the instructions, of the experimental procedures, or of the actual probability distributions faced by individuals. The missing component could, however, have been emotional sink-in that could have resulted from EF's procedures. As detailed in EF's Experimental Methods appendix, these included (a) written instructions on the computer screen, with subjects stepping through each line individually; (b) verbal instructions read by the experimenter to the entire room; and (c) graphical instructions on the computer screen. In contrast, HL's procedures included only written instructions on paper (with no graphics). Our new experiments investigate the idea of the missing sink-in component by adding to HL's experiment a new treatment, designed to directly increase (or not) emotional sink-in.

2.4 New Experiments

Our first experiment was conducted at the RatioLab at the Hebrew University of Jerusalem. The master document from which the experimental instructions were created is available (translated back into English) in appendix A, along with a photo of the experimental setup.[10] The instructions closely follow the instructions from the More Endowment treatment in HL's Experiment 2, with the following adjustments. First, minor adjustments are made to the text for improving flow and streamlining explanations, and to the procedures for adjusting to the different locality. For example, the university-logoed mug and pen are replaced with a reusable water bottle and a chocolate bar. Second, the randomization procedure (involving a random number in a sealed envelope) that determines whether to implement subjects'

[10] As explained in what follows, yellow and purple highlights in the master document each represents a different expectations treatment, while green and red highlights each represents a different demonstration treatment.

keep/exchange choice is simplified: subjects roll a standard (six-sided) die in the end of the experiment; the outcome can either be 4 ($\frac{1}{6}$, or 17% chance) or not ($\frac{5}{6}$, or 83% chance). Thus, in the new experiment, $q \in \{\frac{1}{6}, \frac{5}{6}\}$.

Third, and most importantly, we added an additional instructions page, where participants face a new demonstration-of-probabilities treatment. The new treatment is designed to turn mere (verified) understanding of experiment-relevant probability distributions into something that is more vividly experienced and therefore is more likely to sink in. We first describe the new instructions page that we devised, and then explain the thinking behind its different elements.

Right after participants answer the two comprehension-quiz questions (see p. 3 of the experimental instructions in the appendix) and the experimenter verifies that their answers are indeed correct, and just before filling in the personality test (now on p. 5), the new p. 4 now opens with the text:

> **In this part of the study you will be asked to roll the die a few times, to demonstrate the randomness of a die roll.**
>
> **In front of you are a six-sided die and two markers — one red and one blue.**
> **The die is a fair die: the probability that it will land on any one of its sides is one in six ($\frac{1}{6}$).**
> **Please roll the die. If the result is 4, please color in blue the leftmost square below. If the result is not 4, please color in red the leftmost square below.**
>
> 
>
> **Now repeat this process 17 more times (overall 18 die rolls). After each roll, please color the leftmost empty square in the appropriate color. After 18 rolls, all squares should be colored.**

Half of the participants (randomly assigned within each of the two Expectations treatments) see the above version of the text. The other half see an identical version except that "4" is replaced with "an even number" and "not 4" is replaced with "an odd number." Together with the rest of the new instructions page—which, as described below, is identical across the two versions—we label the versions "1-in-6 Demo" treatment and "50-50 Demo" treatment, respectively. This new manipulation hence extends the original 2×2 experimental design to a $2 \times 2 \times 2$ design: Assignment (bottle vs. chocolate) $\times$ Expectations (weak vs.

strong) × Demonstration (1-in-6 vs. 50-50). The new 1-in-6 Demo treatment is designed to help probability beliefs sink in, by generating a stronger link from participants' probability beliefs—i.e., their mere *understanding* that trade will be possible with probability 1-in-6 or 5-in-6—to actual, repeated and tangible *experience* of probabilistic outcomes; and the new 50-50 Demo treatment is designed to be a clean control: it applies the exact same procedures to a "neutral" distribution that is less relevant to the choice elicited in this experiment.

In the rest of the instructions page participants are asked how many squares they colored in red and blue; they are reminded how the outcome of a later die roll (4 vs. not 4) will determine which item they take home; and they are asked whether the ratio between blue squares and red squares is larger than, smaller than, or equal to the ratio between the chance that their choice will determine which item they take home and the chance that their choice will have no effect. The page closes with the following instructions and questions:

Please answer the following questions. Notice that they probe your subjective feelings regarding probabilities, using verbal expressions.

Next to each question, write down a letter that expresses your feeling regarding the chance that the relevant event will take place, according to this table:

No chance at all	Nearly no chance	Very low chance	Low chance	Slightly low chance	Neither high nor low chance	Slightly high chance	High chance	Very high chance	Nearly certain	Completely certain
A	B	C	D	E	F	G	H	I	J	K

What is the chance that the outcome of the die will be 4? ___. (Write a letter from A to K)

What is the chance that the outcome of the die will not be 4? ___. (Write a letter from A to K)

If I choose "Keep," what is the chance that I will take home the bottle?___. (Write a letter from A to K)

If I choose "Keep," what is the chance that I will take home the chocolate?___. (Write a letter from A to K)

If I choose "Exchange," what is the chance that I will take home the bottle?___. (Write a letter from A to K)

If I choose "Exchange," what is the chance that I will take home the chocolate?___. (Write a letter from A to K)

As repeatedly mentioned above, the purpose of this new page-long Demonstration treatment is to go some way in turning mere beliefs over distributions—those beliefs are verified on the previous instructions page—into something that feels more real to participants, sinks in more, and may hence have more effect on participants' reference point—and therefore, on their choice behavior. In the absence of a theory to guide us how to turn beliefs over

probabilities into those hoped-for choice-relevant reference points, we looked for ideas outside of economics. We found inspiration in the literature on teaching probabilities to children. Aimed at helping children to develop intuitions regarding probabilities and to therefore be able to "grasp the essence" of probability problems (see, e.g., Gage, 2012, and the references therein), that literature discusses several classroom-ready procedures that are devised with the purpose of turning abstract probabilities into more concrete concepts. These procedures include, on the teacher side, using multiple representations and, in particular, using whole numbers rather than fractions (natural frequencies rather than probabilities) and, on the student side, building "visual impressions" of how probabilistic events unfold by recording and tallying realizations using pens in different colors or even tangible Lego-like blocks.

The new Demonstration treatment consists of a combination of these procedures. Consider participants under the 1-in-6 treatment. First, they repeatedly engage in the physical act of rolling the die and recording the outcome using visually contrasting colors; they may therefore "train" themselves both in the emotional exercise of hoping for a certain outcome and comparing it with an actual realization and in the experience of glancing over a vivid graphic that conveys the history of realizations so far. Second, they tally the different outcomes (colored squares) and are asked to compare their ratio with the ratio of probabilities with which their choice vs. the initial coin-flip would determine their take-home item; answering this question requires them to spend some time thinking about said realizations graphic in the context of the odds related to their actual take-home gift. Third, participants choose words ranging from "No chance at all" to "Completely certain" for expressing their "subjective feelings regarding probabilities"; the six questions, in which participants construct statements about their own prospects (e.g., "If I choose 'Keep,' ... the chance that I will take home..."), are aimed at further linking the above physical, visual, and emotional experiences with the beliefs over probabilities that were verified on the previous instructions page.

At the same time, participants under the 50-50 treatment also go through the same procedures (rolling, coloring, tallying, etc.). But they are being "trained" in grasping the essence of a simple probability distribution that is not closely related to their probability beliefs pertinent to the decision problem faced in this experiment. Of course, said procedures

could themselves directly affect behavior, including by having some residual effect on sink-in, or merely by extending the time lag between forming expectations and making a choice. The original 2×2 (no-demonstration) design is therefore an alternative, or an additional, control. Indeed, we had no strong priors, prior to collecting the data, on where in the conceptual span between the original design and the new 1-in-6 treatment we should think of the new 50-50 treatment as placed.

More generally, our combination of the above elements into a Demonstration treatment was based to a large extent on our intuition and on feedback from colleagues, rather than on a formal model. Hence, prior to collecting data, we considered our Demonstration treatment an exploration of ideas rather than an ultimate test of some formal model of sink-in, and we acknowledged that a finding that the treatment has no detectable effect on choice or, alternatively, a very large such effect, should be viewed as the result of a joint test of our ideas (including our sink-in hypothesis), intuitions regarding procedures, and—as always—somewhat arbitrary implementation details.

Shortly after our first experiment, we conducted a second experiment at a different facility of the RatioLab. It was aimed as a replication attempt of the new 1-in-6-Demo treatment. Its instructions and procedures closely follow that treatment in the first experiment, with the following arguably modest modifications that, we hoped, would slightly improve the experimental design (an English translation of the master document is available in appendix B, along with a photo of the experimental setup). First, we rearranged the questions at the bottom part of p. 4—the Demonstration treatment page—in a way that was meant to streamline flow and improve readability.[11] We made this change because an analysis of response patterns in the first experiment suggested that the original flow may have been unnecessarily confusing to some subjects.[12] Second, we modified the opening sentence on

[11] Recall, the Demonstration treatment page (which opens with the text reproduced on p. 14 above) asks subjects to roll a die 18 times, color a box blue when the die shows 4 and red otherwise, tally the colored boxes, and answer questions aimed at further linking these physical, visual, and emotional experiences with the probability beliefs that are relevant to the experiment. Our rearranging of the questions consisted of eliminating a question about probability ratio, clarifying a reminder about experimental procedures and moving it closer to the relevant questions, and boldfacing "**keep**" and "**exchange**" to improve readability.

[12] We report this analysis in appendix G, and show that the revised flow in Study 2 indeed seems to improve on the original flow, with little to no evidence suggesting confusion regarding the demonstration-page questions in Study 2. (When reporting Study 1's results below, we also report them for the subsample of subjects that excludes those whose response patterns may suggest confusion.)

p. 6—the choice-elicitation page—from "You will shortly roll a die to determine whether you can choose to exchange the item you own." to "You will shortly roll a die to determine whether your choice, which you will make on this page, will determine which item you will take home with you." This modification was aimed at reducing the possibility of a demand effect due to potentially implying that choosing "exchange" is desirable. Third, we made a few modifications to the rest of the instructions *after* the choice-elicitation page; these modifications could therefore not affect choice.[13]

2.5 Results

269 subjects participated in the first experiment (Study 1). It ran from March 28 to May 18, 2016, in 19 sessions with estimated average and median duration of 38 minutes.[14] An additional 211 subjects participated in our follow-up, second experiment (Study 2), which was conducted from May 26 to June 16, 2016, in 15 sessions with average and median duration of 38 and 37 minutes, respectively.[15] Of the 269 and 211 participants in the two experiments, respectively, 246 (91.4%) and 196 (92.9%) answered the two comprehension questions correctly on first attempt. Our results are based on all participants; as a robustness check, we also report results for only these correct-on-first-attempt participants.

Table 1 presents our main results. The table's structure is a condensed version of the structure of HL's tables, which in turn are modeled after Plott and Zelier's (2007) Table 1.

Panel A summarizes the findings from our two new experiments. We start with Study 1's 50-50 Demo treatment. Its Weak Expectations column reports an endowment effect of 22 percentage points, driven by coin-flip assignment: while 74% of the 34 coin-bottle participants chose to keep the bottle, only 52% of the 31 coin-chocolate participants chose to exchange the chocolate for a bottle. The effect is sizable but not statistically strong (two

[13]Most notably, we added new post-choice statements to the exit questionnaire on p. 7 of the instructions (see appendix B). These were meat to explore new ideas for future research.

[14]Exact start/end time was not recorded in six of the sessions. Each session included 12–16 participants, seated in cubicles in one or two lab rooms (with 4–12 participants per room). Within each room, the four Expectations×Demonstration treatments were preassigned to alternating cubicles (unbeknown to participants). 40% of the participants were male.

[15]Each session included 8–19 participants, seated in cubicles in one lab room (located on a different campus relative to the first experiment); the two expectations treatments were preassigned to alternating cubicles. Excluding 4 participants who entered conflicting gender data, 57% of the participants were male.

sample, two-sided equality-of-proportions test p-value = 0.07). Next, the Strong Expectations column of Study 1's 50-50 Demo treatment shows that once coin-flip assignment determines final consumption with 5-in-6 rather than 1-in-6 probability, said endowment effect shrinks to 0 ($p = 0.98$), as 62% of both coin-bottle and coin-chocolate participants choose the bottle. While we cannot statistically reject equality of the effect across the two expectations treatments, we note that the direction of change of the effect (from 22% to 0%) is inconsistent with a PE (or a PPE).[16]

This finding from our 50-50 Demo treatments qualitatively replicates HL's results from the comparable no-demonstration treatments on which our experimental design is based. Panel B summarizes these results (reproduced from HL's tables 2 and 3). In the comparable treatment in HL's Study 2 an endowment effect of 31 percentage points ($p = 0.01$) shrinks to 11 points ($p = 0.40$), and in HL's Study 3 an effect of 16 points ($p = 0.08$) shrinks to 8 points ($p = 0.40$).[17] While we again lack the power to reject equality of the effect across the Weak and Strong treatments, the picture that emerges across the three experiments is rather consistent. Taken together, the data are suggestive of a mechanism that affects choice in a direction weakly opposite to KR's PE and PPE predictions. Of course, the KR mechanism may also be at play; however, our results suggest that in HL's no-demonstration setup and in our 50-50 Demo setup, that opposite mechanism—whatever it may be—has a consistently stronger effect.

The picture changes dramatically when switching from 50-50 Demo (or no demonstration) to 1-in-6 Demo, as shown in the rightmost two columns of panel A, Study 1. Now, under Weak Expectations there is no endowment effect—indeed, the table shows an insignificant "anti-endowment" difference of −6 percentage points ($p = 0.60$)—while under Strong

[16] As summarized in the last paragraph of section 2.2 above, the model predicts an endowment effect at low q's (Strong Expectations) regardless of which equilibrium concept, PE or PPE, applies. In contrast, we find no endowment effect (0%). At high q's (Weak Expectations), the effect is never predicted to grow from its low-q level; it could remain constant, shrink, or reverse—depending on how subjects choose among multiple PE. In contrast, we find a large but not statistically strong effect (22%). Finally, if subjects choose a PPE, the model predicts the low-q effect—which we do not find—to remain constant as q increases, and eventually shrink once q exceeds a certain cutoff that depends on η and λ and may lie above our Low Expectations $q = \frac{5}{6}$.

[17] As discussed above and as reported by the q's in Table 1, Weak and Strong Expectations in HL's experiments are somewhat weaker and stronger, respectively, than ours. However, the differences—1-in-10 and 9-in-10 probability of coin-flip assignment determining final consumption, relative to 1-in-6 and 5-in-6 in our experiments—may be too small to be detectable in our data.

Table 1: Choice by Experimental Condition (All "More Endowment" Treatments)

	Weak Expectations	Strong Expectations	Weak Expectations	Strong Expectations
A. New Experimental Treatments	50-50 Demonstration		1-in-6 Demonstration	
Study 1 (N = 269)	q = 5/6	q = 1/6	q = 5/6	q = 1/6
(# coin-bottle, # coin-chocolate)	(34, 31)	(42, 26)	(28, 40)	(38, 30)
(% choosing bottle, % choosing bottle)	(74%, 52%)	(62%, 62%)	(54%, 60%)	(74%, 53%)
difference in proportions	d = 22%	d = 0%	d = –6%	d = 20%
p-value	p = 0.07	p = 0.98	p = 0.60	p = 0.08
Study 2 (N = 211)			q = 5/6	q = 1/6
(# coin-bottle, # coin-chocolate)			(58, 48)	(57, 48)
(% choosing bottle, % choosing bottle)			(59%, 50%)	(61%, 42%)
			d = 9%	d = 20%
			p = 0.37	p = 0.04
B. Treatments in Heffetz & List (2014) (HL)	No Probability Demonst.			
HL Study 2 (N = 117, More Endowment)	q = 9/10	q = 1/10		
(# coin-mug, # coin-pen)	(32, 26)	(25, 34)		
(% choosing mug, % choosing mug)	(81%, 50%)	(64%, 53%)		
	d = 31%	d = 11%		
	p = 0.01	p = 0.40		
HL Study 3 (N = 225)	q = 9/10	q = 1/10		
(# coin-mug, # coin-pen)	(56, 61)	(44, 64)		
(% choosing mug, % choosing mug)	(70%, 54%)	(61%, 53%)		
	d = 16%	d = 8%		
	p = 0.08	p = 0.40		
C. Pooling All Treatments from A + B	50-50 or No Demonst.		1-in-6 Demonstration	
(N = 822)	(122, 118)	(111, 124)	(86, 88)	(95, 78)
	(74%, 53%)	(62%, 55%)	(57%, 55%)	(66%, 46%)
	d = 21%	d = 7%	d = 2%	d = 20%
	p = 0.001	p = 0.26	p = 0.75	p = 0.008

Notes: Percentages are rounded to the nearest whole number. All p-values are from two-sample two-sided tests of equality of proportions. Results in panel B are reproduced from Heffetz and List (2014, tables 2 and 3). A two-sided, eight-sample test of equality of diff-in-diff in proportions—testing the null of no effect of the 1-in-6 Demo treatment—yields p = 0.04 in panel A, Study 1, and p = 0.02 in panel C.

Expectations there is a 20-point effect (p = 0.08). This 20-point effect at low q, as well as the direction of its change relative to the higher q, are consistent with PE and PPE. The null hypothesis that this change, from −6 to 20 points, under 1-in-6 Demo is equal to the opposite change, from 22 to 0 points, under 50-50 Demo, is rejected with p = 0.04 (two-sided, eight-sample test of equality of difference-in-difference in proportions, reported in table notes).[18] [19]

Study 2, aimed mainly to collect more data under 1-in-6 Demo (with a slightly improved design), qualitatively replicates the relevant findings in Study 1, albeit less dramatically: a statistically insignificant 9-point effect (p = 0.37) increases to a statistically suggestive 20-point effect (p = 0.04) when moving from Weak to Strong Expectations—consistent with PE and PPE.

Panel C pools together all 822 observations from panels A and B. It thus summarizes all the evidence we have to date from this experimental design, consolidating the 50-50 Demo and no-demonstration treatments.

In the pooled data, under 50-50 Demo or no demonstration, the endowment effect decreases from a statistically significant (p = 0.001) 21 percentage points in Weak to a not significant (p = 0.26) 7 points in Strong Expectations, offering no support to KR's predictions. In contrast, under 1-in-6 Demo, the effect increases from 2 points (p = 0.75) in Weak to 20 points (p = 0.008) in Strong Expectations, consistent with PE and, to some extent and under certain parameter values, with PPE. The null that this 18-point increase is equal to the 14-point decrease under 50-50 Demo or no demonstration is rejected with p = 0.02 (see table notes).

In summary, the evidence in table 1 suggests that while probability beliefs with no (or no *relevant*) probability demonstration appear to impact the endowment effect in a way that

[18] We use a Z-test. Under the null of equality of difference-in-difference in proportions, denoting each sample's proportion and size p_i and n_i, the point estimate $[(p_1 - p_2) - (p_3 - p_4)] - [(p_5 - p_6) - (p_7 - p_8)]$ is approximately normally distributed with mean 0 and standard deviation $\sqrt{\Sigma_{i=1}^{8} p_i(1-p_i)/n_i}$.

[19] Throughout the paper, when discussing statistical tests, we always report p-values. When using adjectives to describe them, we follow the proposal in Benjamin et al. (2017) to adopt new norms that change the conventional threshold for statistical significance. Accordingly, we reserve the term "statistically significant" for results with p-values below 0.005, and refer as "statistically suggestive" to results with $0.005 < p < 0.05$. Our rejection, with p = 0.04, of the null that our demonstration-of-probabilities variation in Study 1 has no effect is, therefore, statistically suggestive.

does not support KR's predictions, probability beliefs with the relevant demonstration—devised to increase sink-in—can reverse this finding, and impact the effect as the model predicts.[20] Appendix tables A.1 and A.2 probe the robustness of this finding by replicating table 1 for two subsets of its 822 subjects: the 647 subjects excluding those who answered at least one of our and HL's probability-understanding quiz questions incorrectly on first attempt; and the 682 subjects excluding those whose response patterns on the probability-demonstration page may suggest confusion (see footnote 12 and the text around it). Results in these tables move around but remain generally similar; statistical significance drops due to smaller samples. In particular, in panel C, the effect of our new 1-in-6 Demo treatment discussed above—an 18-point increase vs. 14-point decrease in table 1 ($p = 0.02$)—becomes an 13-point increase vs. 18-point decrease in table A.1 ($p = 0.04$), and a 20-point increase vs. 13-point decrease in table A.2 ($p = 0.03$).

3 Effort Provision (Table-Counting Paradigm)

3.1 Theoretical Predictions

Consider a different 2-dimensional consumption bundle, $\boldsymbol{c} = (x, e)$, where x is money (a good) and e is effort (a bad). A lab subject in AFGH's (Abeler, Falk, Goette, and Huffman, 2011) and GGSZ's (Gneezy, Goette, Sprenger, and Zimmermann, 2017) experiments provides effort working on a tedious and repetitive task. Her decision when to stop working is the main outcome of interest. She is informed at the outset that once she stops, with probability $\frac{1}{2}$ she will receive her accumulated earnings (we) according to a known piece rate (or wage) w, and with probability $\frac{1}{2}$ she will instead receive a known (possibly degenerate) lottery

[20] The results in the pooled data in panel C under 1-in-6 Demo have another interesting implication. An early purist interpretation of KR suggested that an appropriately minimalist notion of endowment—that strips endowment of the expectations it typically comes with, and, importantly, of past confounds (e.g., à la Plott and Zeiler 2007)—should not in itself generate a detectable endowment effect. As evidence to the contrary has accumulated (e.g., by HL), this idea has mostly been abandoned. These new results may revive a *sink-in-adjusted* variant of this idea. Specifically, that in panel C under 1-in-6 Demo, coin-flip assignment is not found to cause *any* meaningful endowment effect ($d = 2\%, p = 0.75$) in Weak Expectations but generates an effect ($d = 20\%, p = 0.008$) in Strong Expectations, is perfectly consistent with the following updated idea: an appropriately minimalist notion of endowment—that strips endowment of the *appropriately sunk-in* expectations it typically comes with, and of other past confounds—does not in itself generate a detectable endowment effect.

that is independent of her effort. Specifically, she is informed that she will receive: we with probability $\frac{1}{2}$; a relatively high amount H with probability $p \leq \frac{1}{2}$; and a relatively low amount L with probability $\frac{1}{2} - p$.

A subject hence knows early on that she will later have to choose e and will receive as a result the lottery $(we, \frac{1}{2}; H, p; L, \frac{1}{2} - p)$. AFGH and GGSZ vary H, L, or p across treatments (we provide full details in the next subsection), and investigate subjects' effort choices. They derive theoretical predictions under the assumptions that $m_x(x) = x$ and that $m_e(e) = -c(e), c'(e) > 0, c''(e) > 0$. That is, utility increases linearly in money, and decreases in effort according to a cost function $c(\cdot)$ with standard properties.

A full analysis of a subject's maximization problem is included in appendix F. It is substantially more complicated than in the endowment-effect context above. There, finding all possible PE required examining the only two possible choice plans—keep or exchange—and, for each, considering only a single possible deviation. Similarly, finding a PPE involved comparing at most two possible PE. In sharp contrast, in this effort-provision application there is a continuum of possible plans e, and, for each, a continuum of possible deviations e'. The appendix uses two alternative approaches. First, using (rather cumbersome) algebra, it provides a full case-by-case analysis. Then, in its last section (F.4), it shows an intuitive graphical alternative.

Two conclusions emerge from the analysis in the appendix. First, without characterizing how subjects choose among PE, one could not draw conclusions regarding how a subject's chosen effort level changes with p. There is a range—formally, a closed interval—of effort levels e that are PE at a given p, and while the PE range generally moves up with p, it can do so sufficiently slowly that much of the PE range overlaps across all possible p. That is, subjects in a low-p experimental condition may exert more, less, or equal effort relative to subjects in a high-p condition.

Second, if subjects choose a PPE—which we show to be unique—then effort level increases with p for $we \in [L, H]$, and is independent of p otherwise. Intuitively, since consumption utility is linear in money—formally, $m_x(x) = x$—and since gain-loss utility is piecewise linear—formally, $\mu(x) = \eta x$ for $x > 0$, and $\mu(x) = \eta\lambda x$ for $x \leq 0$—the marginal benefit of effort depends only on the cumulative probabilities of those among the reference points L

and H that are below versus above we. (Without these assumptions, the marginal benefit of effort would in general depend on both the probability and the distance from we of each reference point.) These cumulative probabilities are affected by p only when $L \leq we \leq H$; otherwise, L and H are either both above we, or both below it, with constant cumulative probability ($= \frac{1}{2}$) regardless of p.[21]

3.2 Past Evidence

In AFGH's experiments, a subject provides effort by counting the number of zeros in tables that consist of 150 randomly ordered zeros and ones. AFGH choose this task because (a) it is boring, so effort can be assumed costly; (b) it is pointless (the output is clearly worthless to the experimenter), so reciprocity towards the experimenter can be assumed away; (c) it does not require prior knowledge and offers limited learning possibilities; and (d) performance in it is easily measurable. To familiarize subjects with the task (so they learn their cost of effort $c(\cdot)$ prior to forming expectations), as well as to get a measure of subjects' performance, the experiment contains a first, preparatory part, where subjects are introduced to the task and work on it for four minutes for a piece rate of 10 cents per completed table.

Having completed the first part, subjects receive instructions for the second, main part of the experiment, introducing the experimental setup from the previous subsection. The piece rate is $w = 20$ cents per correctly counted table, the lotteries are degenerate—$p \in \{0, \frac{1}{2}\}$—and the fixed payments are $L = 3$€ and $H = 7$€. Thus, a subject knows that once she stops counting tables, with 50% chance she will receive her acquired earnings, and with 50%

[21] In the theoretical-predictions sections of their papers, AFGH and GGSZ derive predictions using a different equilibrium concept: *choice-acclimating personal equilibrium* (CPE). Introduced in Kőszegi and Rabin (2007), CPE is meant to capture situations where one commits to a choice well in advance ("makes a committed decision long before outcomes occur"), as in, e.g., the purchase (or not) of an insurance policy. Since early commitment rules out the possibility of later deviation, one chooses her preferred option among *any* option in the choice set—taking into account the lagged expectations that her early choice will induce—rather than choosing a PPE among only the (more limited, and more difficult to find) set of PE.

The analysis of CPE is, therefore, substantially more straightforward than that of PE and PPE. However, CPE appears less relevant in this setup because subjects cannot commit well in advance to a certain amount of effort. For this reason, we derive predictions from PE and PPE (as we do for the endowment-effect application). We note however that a CPE analysis leads to the same qualitative conclusions above from a PPE analysis: for we outside the interval $[L, H]$, effort level is independent of p, while inside the interval it increases with p. (See GGSZ's detailed appendix for some discussion, and analysis of PE and PPE, along with additional equilibrium concepts and alternative expectations-based reference-dependence models, e.g., Bell 1985, and Loomes and Sugden 1986.)

chance she will instead receive a known fixed payment—3€ if she is in a "low" treatment; 7€ if she is in a "high" treatment. Importantly, because the lottery is degenerate, and because subjects do not know that the other treatment exists, subjects in the 3€ condition do not imagine that they could have been in a 7€ condition, and vice versa.

As has been noted in the literature, this design is confounded: it could make the fixed payment a natural attractor in the eyes of subjects, resulting in more effort in the 7€ condition (and more bunching around 7€) independent of any KR-predicted effects. For example, as Ericson and Fuster (2011) note in their introduction, AFGH's effort-provision experiment is subject to the same confound that was found in EF's own endowment-effect experiment to strongly affect (indeed, reverse) outcomes: "In this study, subjects are not informed that the level of the fixed payment (high/low) is randomly assigned. As a consequence, subjects may make (potentially mistaken) inferences from the amount of the fixed payment about the "appropriate" amount of effort to provide, and this could contribute to the observed treatment effect (e.g., if the fixed payment is perceived as an indication of how much the experimenter expects participants to work)."[22] While in EF's experiment the confound of subjects' potential inferences is expected—and was indeed found—to move behavior in a direction *opposite* to that predicted by the KR model, in AFGH's experiment this confound is expected to move behavior in the *same* direction as that predicted by the model.[23]

AFGH's main finding is that on average, subjects (N = 120) work for a longer time (p = 0.044) and earn more money (p = 0.046) in the high (7€) treatment. The general direction of the high-low difference is consistent with a PPE. But subsequent analysis and data collection suggest that overall, the evidence is less supportive of the model than it first appeared. We highlight four points. First, inconsistent with PPE (or, for that matter, CPE; see footnote 21), some of the high-low difference is found *outside* the $[3, 7]$ interval.

[22] In an earlier draft, Ericson and Fuster (2010) add a concrete example: subjects could, for instance, "think that the fixed payment they may receive instead of their piece rate earnings was calibrated to be close to the piece rate earnings the average participant accumulates."

[23] AFGH take great care in their thoughtfully designed baseline experiment to rule out almost all *other* alternative mechanisms and confounds one could think of. For example, they convincingly rule out peer effects, social comparisons, and conformity considerations, by having subjects arrive to the experiment one at a time (at least 20 minutes apart) and be alone in a room during the experiment. They also rule out, by running additional clever treatments, gift-exchange motives and the worry that the fixed payment is particularly salient.

Second, the main finding, i.e., the high-low difference in average earnings, did not strongly replicate with a larger sample ($N = 318$) by Camerer et al. (2016), who found an insignificant ($p = 0.160$) effect with relative size 36% of the original. Third, as discussed above, the general direction of the difference is also predicted by simpler explanations, some of which could even be considered simple "demand effect" explanations. (Moreover, such alternative explanations may be more consistent with a difference inside as well as outside the $[3, 7]$ interval.) Finally, in what is arguably a cleaner design—three of GGSZ's nine treatments discussed below—findings are inconsistent with PPE. We now elaborate on this last point.

In the conclusion of their paper, AFGH propose for future work a version of their experiment that replaces the single fixed payment with a (nondegenerate) lottery across two distinct fixed payments. Formally, their proposal amounts to switching from $p \in \{0, \frac{1}{2}\}$ to $p \in (0, \frac{1}{2})$. The modified experiment is designed to distinguish between models where the reference point is the mean of the expected outcomes (e.g., Bell 1985, Looms and Sugden 1986, Gul 1991) and models where the reference point is the entire distribution of expected outcomes (e.g., KR). We note that the modified experiment is, in addition, a cleaner test of these models: since a nondegenerate lottery means that all subjects are presented with both L and H and may receive each with a positive probability (which differs across treatments), any mistaken inferences from L or from H to the "appropriate" amount of effort to provide are likely to affect subjects much more uniformly across treatments.

GGSZ run nine different treatments. In addition to the original two treatments, which across the Atlantic become \$3 and \$7, they run seven new treatments, including three that implement the proposed (nondegenerate) lottery design. In these three treatments, $L = \$0$, $H = \$14$, and $p = \frac{1}{8}$, $\frac{2}{8}$, or $\frac{3}{8}$. Any mistaken inferences by subjects from L and H are held fixed across all subjects: in all treatments subjects are exposed to both L and H, as they may actually receive both \$0 and \$14, each with at least $\frac{1}{8}$ probability. In addition, as noted by GGSZ, the extreme levels of L and H may make them less likely to be perceived as informative signals regarding the appropriate amount of earnings.

GGSZ's findings in these three treatments are difficult to explain by any model we are aware of: effort sharply *decreases* from $p = \frac{1}{8}$ to $p = \frac{2}{8}$, and then recovers in $p = \frac{3}{8}$ to a little

above its level in $p = \frac{1}{8}$.[24]

Overall, our interpretation of the data from AFGH's and GGSZ's effort-provision experiments in this table-counting paradigm is that they do not provide clean or robust evidence consistent with a PPE.[25] Trivially, the data are consistent with PE: without specifying how subjects select among multiple PE, the model does not rule out change in either direction (or no change) of effort with p.

We summarize our reading of this part of the literature as follows. AFGH's baseline design is confounded in a way that has been shown by EF (albeit in a different setting) to dramatically affect behavior. In the cleaner, nondegenerate-lottery design of GGSZ, where, from subjects' perspective, all is held fixed other than the relative probabilities of the two fixed payments, the evidence contrasts with any model we are aware of—expectations-based or not. Moreover, while confounded and therefore difficult to interpret, AFGH's baseline findings were found substantially smaller in Camerer et al.'s (2016) replication attempt with a substantially larger sample. (Also, GGSZ's extension of AFGH's treatments to additional fixed-payment treatments show that they do not easily generalize; see footnote 24.)

Our new experiments add sink-in manipulations to two new versions (high and low) of GGSZ's nondegenerate-lottery treatments. Their purpose is to investigate whether the absence of clean and robust PPE(/CPE)-consistent evidence in the table-counting paradigm could be explained, at least in part, by expectations not having appropriately sunk in. As we report below, the answer appears to be negative: our new treatments do not significantly change the above picture, as we fail to find significant high-low differences in effort. In our concluding discussion in section 4 we discuss other potential reasons for this general absence of clean and robust PPE(/CPE)-consistent evidence in this paradigm.

[24] We reproduce these findings below (section 3.4, table 2, panel B). GGSZ's other treatments effectively repeat AFGH's original single-fixed-payment design, with \$0, \$3, \$7, or \$14, and are therefore subject to the original confound. The findings in these treatments, like those in the three cleaner treatments we focus on, are also difficult to explain: effort is lower in the \$3 than in the \$7 treatment (as in the original experiment), but in the \$0 treatment and \$14 treatment effort is roughly as high as in the \$7 treatment. See GGSZ's figure 2.

[25] Equally, they do not provide clean or robust evidence consistent with a CPE (see footnote 21). Of course, the fact that GGSZ, in their cleaner (nondegenerate) implementation of this paradigm, find that subjects respond to the probability p in *any* direction provides strong evidence against the standard model, where p should not affect behavior.

3.3 New Experiments

Our first effort-provision experiment (Study 3) was conducted at the Business Simulation Lab at Cornell University. The master document from which the experimental instructions were created is available in appendix C, along with a photo of the experimental setup.[26] As explained in the previous subsection, the instructions come in two parts. The instructions closely follow GGSZ's instructions, which in turn closely follow (the English translation of) AFGH's instructions. We made the following adjustments. First, we increased the show-up fee from \$5 to \$10 (it was 5€ in AFGH). Second, we again replaced the original randomization procedure (involving cards in two or eight sealed envelopes) with a procedure involving a coin flip and a die roll: when a subject decides to stop working, she flips a coin; if she flips heads, she receives her acquired earnings; if she flips tails, she rolls a die to determine whether she receives \$0 or \$14, according to whether she rolls a 4 or not, and to which experimental condition she is in. While theory invariant, this modification serves three purposes: (a) it makes the Demonstration treatment from our endowment-effect experiments readily importable to this setup; (b) it slightly expands the probability variation across treatments, to $\frac{1}{12}$ vs. $\frac{5}{12}$ (from $\frac{1}{8}$ vs. $\frac{3}{8}$) without having to use many envelopes; and (c) it may help to shift subjects' focus from this unconditional probability variation to the *conditional* probability variation 1-in-6 vs. 5-in-6 (conditional on flipping tails), which, due to being twice as wide, may help make the variation feel more substantial.

Most importantly, as in our first two experiments, we added an additional instructions page, containing the Demonstration treatment. The additional page follows what used to be the last instructions page in GGSZ. Thus, just after participants answer the example questions and the experimenter verifies that their answers are indeed correct, and just before they start the second part of the experiment on the computer, the new p. 6 opens with the text: "You are almost ready to start the table-solving task on the screen. Before you do, we would like to demonstrate to you the randomness of a die roll." The rest of the page follows closely the Demonstration-treatment page from Studies 1 and 2, with much of the

[26] Cyan and purple highlights in the master document represent different expectations treatments, while green and red highlights represent (as before) different demonstration treatments. The top and bottom screenshots on page 3 of the document correspond, respectively, with the cyan and purple treatments.

text identical, verbatim, to the text in Study 2 (see footnote 11 above).

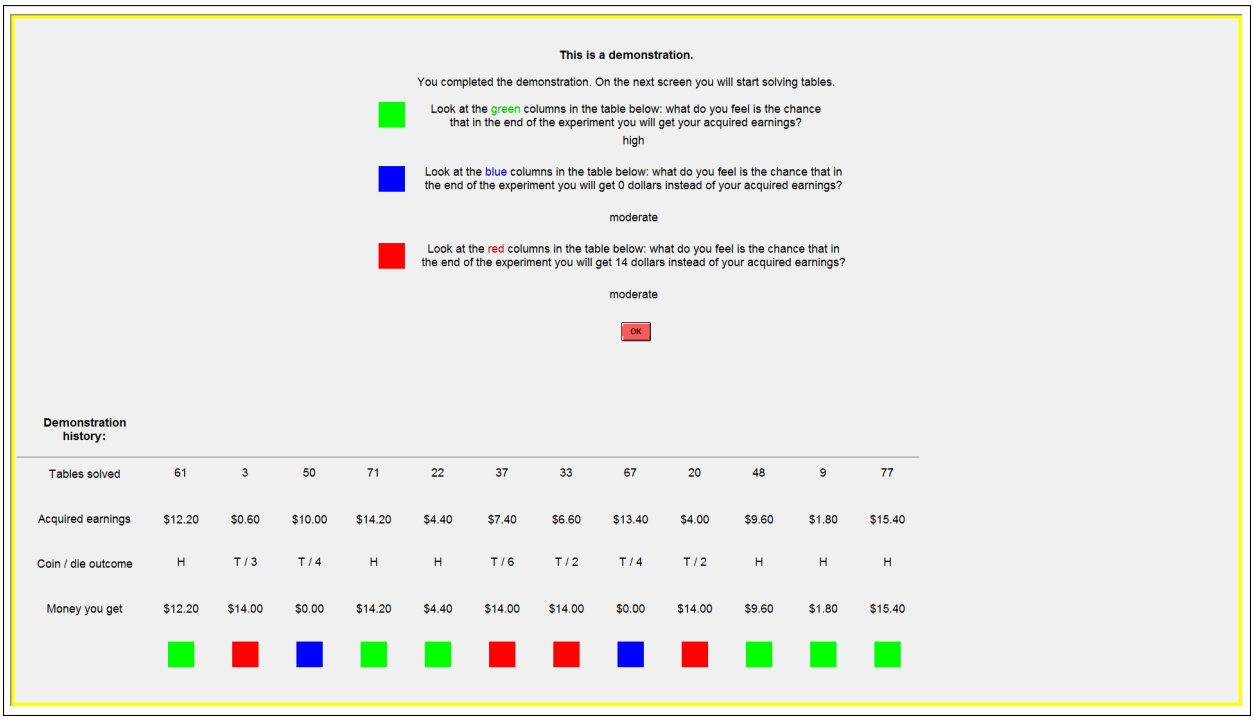
Having completed the Demonstration page, participants start the second part of the experiment on the computer screen. The experiment is implemented using the z-Tree (Fischbacher 2007) code generously provided by GGSZ. We made no substantial modifications to the code besides adjusting the wording of the on-screen reminder (of the randomization procedure and condition) to match our modified randomization procedure (see screenshot on p. 3 of the instructions in the appendix).[27]

Study 4 was conducted three months after Study 3. Like Study 2, its main purpose was to collect more data under a 1-in-6 Demo condition. However, in contrast with Study 2, whose design closely followed that of Study 1 in order to probe the replicability of Study 1's findings, in Study 4 we aimed to create a yet stronger 1-in-6 Demo treatment in order to stress-test Study 3's *non*-findings. With this motivation, we moved Study 4's probability-demonstration component from the paper instructions to the computer screen. This allowed us to adapt it to the specific randomization procedures in this experiment, creating a demonstration that attempts to closely simulate subjects' coin-and-die experience. Since the setup of Studies 3 and 4 was otherwise identical, Study 3's instructions and photo (provided in appendix C) remain relevant for Study 4, with only three differences. First, the blue and red markers (in the photo) are no longer provided. Second, the probability demonstration page (the last page of the instructions) is dropped. And third, the last sentence of the remaining instructions is modified to read: "After you have answered the example questions correctly, the experimenter will start a computerized demonstration on the screen. After the demonstration, you will start the second part of the experiment."

Appendix D provides screenshots of the computerized demonstration, followed by a detailed step-by-step description of subjects' experience during the demonstration. Here we provide a brief overview. The header "**This is a demonstration.**" appears in bold on all screens. An early screen introduces the die on the table as a fair die (as in Studies 1–3)

[27]While we purposefully kept the instructions language (both on paper and on the screen) identical to GGSZ's wherever possible, we did change "on" to "into" in the original on-screen instruction "Please contact the experimenter by stepping on the corridor" (this line appears only after subjects stopped working, i.e., after all data have been collected). On the other hand, although we noticed that the original z-Tree code in fact considers miscounts of at most one zero (in either direction) as correct counts, we modified neither the code nor the instructions, opting instead to keep this minor discrepancy between code and instructions, thus keeping results as comparable as reasonably possible.

Figure 1: Study 4's Probability Demonstration (Last Screen)



Notes: See appendix D for step-by-step screenshots and detailed explanation.

and, in addition, the coin on the table as a fair coin (unlike previous studies). The screen explains, among other things, that "we would like to demonstrate to you the randomness of coin flips and die rolls," and that during the demonstration

> we will ask you to imagine that you solved different amounts of tables. For each amount, we will ask you to flip the coin, and if you flip tails, we will ask you to also roll the die. In total, we will ask you to do this twelve times, and each time you will enter the coin and die outcomes.

On the following screens, as subjects follow this procedure, a "Demonstration history" table adds one column at a time, each time recording a (randomly chosen) would-be number of tables solved, its corresponding amount of acquired earnings, the outcome of the coin and, if relevant, the die (after subjects actually tossed or rolled them), and the amount of money the subject would hypothetically get under the demonstrated scenario. Different colors mark each column according to that outcome: green for acquired earnings (in both treatments), blue or red for \$0 (by treatment), and red or blue for \$14 (by treatment). The demonstration ends with three questions regarding subjects' felt chance of each outcome. Figure 1 reproduces an example of the final demonstration screen, with all of the subject's answers recorded and displayed. The average duration of Study 4's demonstration was similar to that of Study 3: almost 12 versus almost 13 minutes, from the end of each experiment's first stage to the start of its second stage.[28]

3.4 Results

127 subjects participated in Study 3, which ran from September 15 to October 20, 2016.[29] An additional 120 subjects participated in Study 4, which ran between January 24 and February

[28]For completeness and comparability with Studies 1 and 2, appendix figures A.5–A.8 present, by experimental cell, the means and 95% confidence intervals, and the entire distributions, of responses on the Probability Demonstration page (Study 3) or screens (Study 4). The figures convey no big surprises. They suggest that, first, subjects react as expected to Demonstration condition in terms of fraction of blue squares and, in Study 4, also of green squares. Second, subjects also respond largely as expected to the feelings-regarding-chances questions. Finally, as in Studies 1 and 2, Demonstration condition has essentially no effect on subjects' responses to the feelings questions.

[29]We had to discard 14 prior observations and restart the experiment on September 15, after discovering and fixing a z-Tree programming error that created a mismatch between the experimental condition in the written instructions and on the screen. We had to discard another participant who, in a session on October 12, was mistakenly handed the wrong instructions, creating a similar mismatch. These are not included in our 127-subject sample.

23, 2017. Table 2's panel A presents the two studies' main findings. The six results columns correspond to the four treatments in Study 3 followed by the two treatments in Study 4. For comparability, each column reports the statistics that GGSZ report in their table 1: p, N, mean, median, and S.D. In addition, the panel reports the difference (d) in acquired earnings between each of the three High-Low pairs, and its p-value.

We start with Study 3. Under 50-50 Demo, average acquired earnings of the 31 subjects in Low Expectations ($p = \frac{1}{12}$) is \$5.81, compared with \$6.65 for the 31 subjects in High Expectations ($p = \frac{5}{12}$). The difference, d = \$0.84, is in the direction predicted by KR, but is not statistically significant (t-test $p = 0.42$). For comparison, panel B reports results from the treatments in GGSZ on which our experimental design is based. It shows that the \$0.84 difference we find under 50-50 Demo essentially replicates the (also not significant) \$1.00 difference found by GGSZ with no probability demonstration, between their two treatments that are closest to ours, i.e., $p = \frac{1}{8}$ and $p = \frac{3}{8}$. (For completeness and as a reminder, panel B also shows that in another GGSZ treatment, $p = \frac{2}{8}$, that is between these two treatments, acquired earnings are significantly lower, creating a sharp non-monotonicity that is not predicted by any model we are aware of—including, of course, the standard model—and that does not easily lend itself to a sink-in explanation either.) Back to panel A, the \$0.84 difference between Low and High Expectations does not change much (it grows slightly, to $d^c = 0.98$, with $p^c = 0.33$) in an OLS regression that linearly controls for productivity (measured as acquired earnings in the experiment's first part), hour of the day (10am–7pm), school day vs. weekend/break (0/1), and female vs. male (0/1).

The next four columns in panel A show that our 1-in-6 Demo treatments do not make the High-Low difference (significantly) larger. In fact, in Study 3, the (still insignificant) High-Low difference is now *negative* (d = –\$0.56, $p = 0.70$; d^c = –\$0.97, $p^c = 0.48$). Study 4's computerized demonstration, where subjects experience both the coin and the die, may have been more effective in moving the results in the KR direction: its High-Low difference is positive and is 31–45% larger than Study 3's difference under 50-50 Demo, but it is still statistically insignificant (d = \$1.22, $p = 0.18$; d^c = \$1.28, $p^c = 0.15$).

Finally, for completeness and for further increasing backward comparability, appendix table A.3 reproduces table 2 but replaces acquired earnings with time spent working—an

Table 2: Effort by Experimental Condition (All $0 < p < \frac{1}{2}, H = \$14, L = \$0$ Treatments)

		Low Exp.	High Exp.	Low Exp.	High Exp.	Low Exp.	High Exp.
A. New Experiments		Study 3: No Coin				Study 4: Coin	
		50-50 Demonst.		1-in-6 Demonstration			
	p	1/12	5/12	1/12	5/12	1/12	5/12
	N	31	31	31	34	60	60
Acquired Earnings	mean	5.81	6.65	6.42	5.86	5.14	6.36
	[med]	[4.6]	[5]	[5]	[5]	[4]	[5]
	(S.D.)	(4.52)	(3.64)	(5.71)	(5.95)	(4.37)	(5.47)
		$d = 0.84$		$d = -0.56$		$d = 1.22$	
		$p = 0.42$		$p = 0.70$		$p = 0.18$	
		$d^c = 0.98$		$d^c = -0.97$		$d^c = 1.28$	
		$p^c = 0.33$		$p^c = 0.48$		$p^c = 0.15$	
B. Gneezy et al. (2017)		No Prob. Demo.					
	p	1/8	2/8	3/8			
	N	30	29	29			
Acquired Earnings	mean	7.35	4.86	8.34			
	[med]	[5]	[4]	[9]			
	(S.D.)	(5.12)	(3.85)	(5.38)			
	$d =$	−2.49	3.49				
	$p =$	0.04	0.006				
		⌙—— $d = 1.00$ ——⌝					
		$p = 0.47$					

Notes: Acquired earnings: in dollars. Differences and p-values: OLS regressions on a treatment indicator (and a constant), without additional controls (d, p) and with the following controls (d^c, p^c): productivity (in the experiment's first part), time of the day, and indicators for weekend/break and male. Each regression includes observations from a pair of treatments. Results in panel B are reproduced from Gneezy et al. (2017, table 1) or calculated from a data package provided by the authors.

additional outcome variable that AFGH investigate in their original experiment. Qualitatively, our (non)findings remain the same. In particular, all the signs of the High-Low differences discussed above remain the same, as does their lack of statistical significance.

4 Discussion and Conclusion

All models are wrong. Defining reference points as lagged probability beliefs may be a useful simplification for writing an applicable theory, but in reality reference points may be more than merely lagged beliefs. The evidence presented in this paper supports a variant of this hypothesis in one endowment-effect context. In one effort-provision context, however, where clean, robust evidence of reference points as lagged beliefs has previously not been consistently found, our attempt to replace merely lagged beliefs with sunk-in beliefs does not significantly strengthen the evidence.

Under the sink-in hypothesis, while *clean* tests of the theory need to manipulate reference points by carefully manipulating nothing but lagged probability beliefs, *useful* tests of the theory need to manipulate reference points by manipulating sunk-in beliefs. Yet, the sink-in hypothesis does not rule out the possibility that while a *correct* theory should define reference points as sunk-in beliefs, a *useful* theory should keep defining them as lagged beliefs.[30]

These questions of useful versus clean/correct theories and tests are closely related to

[30]In other words, even if the sink-in hypothesis is correct, it does not necessarily follow that sink-in should be explicitly incorporated into the model. Such rewriting of the model may make it a more complete and accurate explanation of the world, but one could argue that a model should not aim for such completeness and accuracy in the first place. While a comprehensive discussion is beyond the scope of this paper, we note that according to this view, the KR model correctly reduces the rich, intractable, real-world concept of expectations to the simple formal concept of lagged probability beliefs (or information regarding probability distributions). Although in the real world probability beliefs are rarely exogenously endowed and often emerge with experience and learning, over time, in ways that render them inherently different from endowed probability beliefs, designing "clean" lab experiments that disentangle reference-point-relevant expectations from mere probability beliefs to demonstrate said difference is a futile exercise. Of course the model is incomplete—any model is an inherently incomplete rendering of reality—but it is tractable and hence useful. While it may be useful to investigate the links between theory and evidence—and in particular, in our context, to investigate what else, beyond lagged probability beliefs, is necessary to endow subjects with in order to find KR-esque effects in the lab—lagged probability beliefs should be viewed merely as a useful modeling device, and should not have been taken literally in the first place.

In this view, rather than using our findings in this paper to change the theory (i.e., the math, or the definitions), one should use them to clarify the link between theory and evidence, and to make sure that when designing clean tests of the theory, the tests are made clean on some dimensions but not on those where we already know that the theory is not (and is not meant to be) literally true.

another question: what is versus what is not an experimental confound. Consider Knetsch's (1989) first endowment-effect demonstration in his original exchange-paradigm experiment. His subjects overwhelmingly chose to keep a randomly-assigned owned object rather than trade it for another. But his experiment has subsequently been considered confounded, for example due to the information considerations discussed above, demand effects, transaction costs (the endowed item is already in subjects' possession when they have to make a choice), and other reasons. In the context of KR's model, an additional potential design weakness has been pointed out: reference points *as lagged probability beliefs* are not directly manipulated; instead, subjects are simply endowed with an item, and the researcher has to infer their expectations regarding future consumption. But under the hypothesis that expectations have to sink in to become reference points, Knetsch's original design, while not clean, may be useful. Indeed, it may be more useful than the arguably cleaner designs of EF and HL: simply endowing a subject with a mug as a gift may help expectations sink in more effectively than carefully presenting a subject with a mug and laboriously explaining a randomization procedure through which, with high probability, she will keep it no matter what she later chooses. Similarly, in the effort-provision context, what we and others refer to as AFGH's degenerate-lottery confound may have had an effect on choices in part because it is an effective way, if not a clean way, to make expectations sink in. (This account however does not explain the results from the follow-up studies that replicated AFGH's original degenerate-lottery design.)

While we acknowledge these possibilities, our new treatments are designed to isolate, and directly manipulate, a sink-in component. In the endowment-effect setting, where these new treatments are found to have an effect, it is important to also note what they do *not* do. Relative to HL's original no-demonstration findings and their qualitative replication in our 50-50 Demo condition, the 1-in-6 Demo condition does not improve (indeed, change) the explanation of the experimental procedures or of the relevant probabilities, and, measured by the proportion of subjects who correctly answer the quiz questions on first attempt, it does not meaningfully improve subjects' understanding of the procedures. In addition, relative to the 50-50 Demo condition, the 1-in-6 Demo condition does not change the time that passed between the outset of expectations formation and eventual choice. Finally, the

reversal of the direction of change of the endowment effect across these two conditions rules out a probability-weighting explanation.[31]

We close this discussion with speculation regarding the failure of the effort-provision table-counting paradigm to provide clean and robust KR-consistent evidence to date, including under our new sink-in manipulations in Studies 3 and 4, and its contrast with our endowment-effect findings in Studies 1 and 2.

We start with theory. When analyzing the experimental setups and deriving theoretical predictions in sections 2.1, 2.2, 3.1, and appendix F, we followed the original papers (i.e., AFGH, EF, GGSZ, HL) and used the basic framework in Kőszegi and Rabin (2006, 2007). Kőszegi and Rabin (2009) embed this basic framework within a general multiple-period dynamic framework. They generalize earlier equilibrium concepts, including PPE and CPE, and specify the assumptions under which the generalized model and its updated solution concepts reduce to the earlier model and solution concepts, which now become special cases. An analysis of the original experiments using this updated "news utility" framework, while overdue, is beyond the scope of the present paper. Simply put, such analysis would embed the less-than-hour-long experiments within the rest of subjects' lives. In particular, it would take into account the rational expectations, or "consumption plans," subjects hold, upon entering the experimental lab, regarding future consumption—i.e., consumption during the experiment and, importantly, after leaving the lab. It would also analyze the updating of these expectations/plans during the experiment, as subjects learn about new consumption opportunities. Preliminary analysis suggests that the updated framework may affect predictions in the effort-provision setup more than in the endowment-effect setup. One reason is that while subjects likely have relevant expectations, upon entering the lab, regarding the amounts of effort and money they will consume during and after the experiment, they may have effectively no consumption plans regarding small items to be surprisingly endowed with and possibly exchange. We cautiously speculate that a more complete analysis may therefore provide one potential explanation for the hitherto unexplained findings in the table-counting

[31] See, e.g., Hertwig and Erev's "The description-experience gap in risky choice" (2009): "...the general pattern can be summarized as follows: in decisions from experience, people behave as if the rare events have less impact than they deserve according to their objective probabilities, whereas in decisions from description people behave as if the rare events have more impact than they deserve (consistent with cumulative prospect theory)."

paradigm.

Within the special-case framework of KR (2006), as we discuss above and show in appendix F, PE and PPE are substantially more difficult to calculate in the effort-provision than in the endowment-effect experiments. This may itself suggest, again cautiously, that the theory may be less relevant in the specific table-counting setup, where endogenous rational-expectations equilibria are perhaps less likely to instinctively arise. Beyond these cautions arguments, recall that without the PPE refinement and under a wide range of parameter values, PE does not rule out effort that increases, decreases, or remains constant with p. Trivially, this in turn means that without specifying how individuals choose among multiple PE, all past findings (including AFGH, Camerer et al. 2016, GGSZ, and our new Studies 3 and 4) could be accommodated. In contrast, in the endowment-effect setup, the effectively unique PE under low q does imply nontrivial, falsifiable predictions. Specifically, it implies an endowment effect at low q that should not increase with q—consistent with what we find under 1-in-6 Demo in Studies 1 and 2.

In addition to adopting the updated and generalized KR (2009) theoretical framework, and using it to both analyze past evidence and develop new experimental paradigms, future research could investigate further-refined variants of our still-underspecified sink-in hypothesis. For example, related to the above discussion about confounds, consider the notion that the extreme levels of L and H in a \$0/\$14 table-counting design may make them less likely than in a \$3/\$7 design to be perceived as plausible, relevant signals regarding the appropriate amount of earnings. We mentioned this idea in section 3.2 in support of the therefore cleaner \$0/\$14 design. But what if a necessary precondition for expectations to sink in is for them to be perceived as plausible and relevant in the first place? For another example, in an alternative, *competitive* effort-provision paradigm, where a second mover explicitly competes against a first mover, Gill and Prowse (2012) provide evidence supportive of expectations-based reference-dependent-preferences models. They conclude that in their setting, effectively no lag is required for expectations to become reference points: "Given the tiny temporal gap between the agents' effort choices and the outcome of the tournament, our results indicate that, at least in our competitive framework, the adjustment process is essentially instantaneous." Could competition foster sink-in, and could lagged beliefs be replaced

with sunk-in-through-competition beliefs? We leave these questions for future research.

References

Abeler, Johannes, Armin Falk, Lorenz Goette, and David Huffman. 2011. "Reference Points and Effort Provision." *American Economic Review* 101(2), 470–492.

Bell, David E. 1985. "Disappointment in Decision Making under Uncertainty." *Operations Research* 33(1), 1–27.

Benjamin, Daniel J., James O. Berger, Magnus Johannesson, Brian Nosek, Eric-Jan Wagenmakers, ..., Valen Johnson. 2017. "Redefine Statistical Significance." *Nature Human Behaviour*, forthcoming.

Camerer, Colin F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler et al. 2016. "Evaluating replicability of laboratory experiments in economics." *Science* 351(6280), 1433–1436.

Ericson, Keith M. Marzilli and Andreas Fuster. 2010. "Expectations as Endowments: Evidence on Reference-Dependent Preferences from Exchange and Valuation Experiments." *Manuscript*. May 19.

Ericson, Keith M. Marzilli and Andreas Fuster. 2011. "Expectations as Endowments: Evidence on Reference-Dependent Preferences from Exchange and Valuation Experiments." *Quarterly Journal of Economics* 126(4), 1879–1907.

Fischbacher, Urs. 2007. "z-Tree: Zurich toolbox for ready-made economic experiments." *Experimental Economics* 10 (2), 171–178.

Gage, Jenny. 2012. "Towards a New Probability Curriculum for Secondary Schools." *12th International Congress on Mathematical Education*, COEX, Seoul.

Gill, David and Victoria Prowse. 2012. "A Structural Analysis of Disappointment Aversion in a Real Effort Competition." *American Economic Review* 102(1), 469–503.

Gneezy, Uri, Lorenz Goette, Charles Sprenger, and Florian Zimmermann. 2017. "The Limits of Expectations-Based Reference Dependence." *Journal of the European Economic Association*, 15(4), 861–876.

Goette, Lorenz, Annette Harms, and Charles Sprenger. 2017. "Randomizing Endowments: An Experimental Study of Rational Expectations and Reference-Dependent Preferences." *American Economic Journal: Microeconomics*, forthcoming.

Gul, Faruk. 1991. "A Theory of Disappointment Aversion." *Econometrica* 59(3), 667–686.

Heffetz, Ori and John A. List. 2014. "Is the Endowment Effect an Expectations Effect?" *Journal of the European Economic Association*, 12(5), 1396–1422.

Hertwig, Ralph, and Ido Erev. 2009. "The description-experience gap in risky choice." *Trends in cognitive sciences* 13(12), 517–523.

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), 1325–1348.

Kahneman, Daniel, and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica* 48, 263–291.

Knetsch, Jack L. 1989. "The Endowment Effect and Evidence of Nonreversible Indifference Curves." *American Economic Review* 79(5), 1277–1284.

Kőszegi, Botond and Matthew Rabin. 2006. "A Model of Reference-Dependent Preferences." *Quarterly Journal of Economics* 121(4), 1133–1165.

Kőszegi, Botond and Matthew Rabin. 2007. "Reference-Dependent Risk Attitudes." *American Economic Review* 97(4), 1047–1073.

Kőszegi, Botond and Matthew Rabin. 2009. "Reference-Dependent Consumption Plans." *American Economic Review* 99(3), 909–936.

Loomes, Graham and Robert Sugden. 1986. "Disappointment and dynamic consistency in choice under uncertainty." *Review of Economic Studies* 53(2), 271–282.

O'Donoghue, Ted, and Charles Sprenger. 2018. "Reference-Dependent Preferences." In *Handbook of Behavioral Economics*, edited by Doug Bernheim, Stefano DellaVigna, and David Laibson. Elsevier, forthcoming.

Plott, Charles R. and Kathryn Zeiler. 2007. "Exchange Asymmetries Incorrectly Interpreted as Evidence of Endowment Effect Theory and Prospect Theory." *American Economic*

Review 97(4), 1449–1466.

Smith, Alec. 2008. "Lagged Beliefs and Reference-Dependent Utility." University of Arizona working paper #08-03.