THE UNIVERSITY OF CHICAGO

ESSAYS ON INFORMATION IN CONSUMER CONTEXTS

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE UNIVERSITY OF CHICAGO

BOOTH SCHOOL OF BUSINESS

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

BY

SHWETA R. DESIRAJU

CHICAGO, ILLINOIS

AUGUST 2023

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Acknowledgments

I would like to thank my advisors, Berkeley Dietvorst and Oleg Urminsky, my committee members, Abigail Sussman and Dan Bartels, and Pradeep Chintagunta. This work would not have been possible without your support, guidance, and expertise. I have tremendous admiration for you all. I am grateful to Cynthia Hillman and Malaina Brown for shepherding me through the PhD process. And to Drs Madhu and Ram, I am so fortunate to have you as parents.

Overview

In this dissertation research, I explore the influence of information on consumer decision making in two essays. Consumer environments are often set so that individuals are positioned to encode information in a particular way. For instance, a choice architect may set a default and limit accompanying details of other options to encourage individuals to select the default. In another case, a firm may present a product as new by prominently displaying a high version number instead of listing its novel features. In this dissertation, essay 1 explores whether the addition of information atypical to choice architecture can increase preference consistency and essay 2 explores the possible influence of information about product updates on product choice.

However, the influence of information is not plain across settings. Research on choice architecture suggests that individuals will eschew effortful information seeking (Jachimowicz, Duncan, Weber, & Johnson, 2019). Additionally, when information about transparency is incorporated into a default structure, it is not seen to reduce choice of the default option (Steffel, Williams, Pogacar, 2016). Essay 1 demonstrates that consumers can take up additional information when faced with a default, and that doing so improves preference consistency and the decision experience. Research on heuristic cues suggests that individuals will be attracted to larger values (Chinander and Schweitzer, 2003) —as may be seen in products that are the most updated--and work on product updates focuses on existing consumers rather than new users (Okada, 2006, Wood and Lynch, 2002; Jung, Peck, Palmeira, and Kim, 2022). Essay 2 is an exploration into the responses of new consumers and demonstrates that consumer evaluations of product updates are reference-dependent.

In essay 1 (status: forthcoming at *Psychological Science*), I study how including information about preference heterogeneity alongside default choices can help consumers with

different interests make preference consistent choices. In three studies, the efficacy of a default including information about preference heterogeneity is compared to that of standard defaults and forced choices. Including information about preference heterogeneity alongside default choice architecture helps individuals with different preferences sort themselves into the option consistent with their preferences more so than the other common choice architectures tested. Additionally, including this information made consumers feel a choice was more transparent without feeling more difficulty or confusion.

In essay 2 (status: in preparation for submission), I study how information about how often products are updated influences consumer preference for updates. In four studies, consumers appear to utilize information about update frequency as a reference point. Consumers display limited interest in the most updated option. Interest in the more updated option is fostered when an external threshold for update frequency is higher than one's options and decreases as the threshold nears or falls beneath the update frequency of one's options. When an external threshold is not present, a similar pattern is observed based on the personal expectations held by consumers. Additionally, consumers appear to utilize update frequency as an attribute rather than as a signal of quality or production effort.

ESSAY 1

Reason defaults: Presenting defaults with reasons for choosing each option helps decision makers with minority interests

Abstract: Defaults are powerful tools for nudging individuals towards potentially beneficial options. However, defaults typically guide all decision-makers towards the same option, and as a result, may misguide individuals with minority interests. We test whether presenting defaults with information about heterogeneity can help individuals with minority interests select alternative options, and dub this intervention a "reason default." Reason defaults pre-select the option that is best for most individuals (like standard defaults), but also explain 1) why the default was selected and 2) who should opt for an alternative. In five preregistered studies using online convenience samples (N=4,210) we find that reason defaults can improve decision-makers' outcomes over standard defaults and forced choices by guiding most individuals towards the default option, while helping individuals with minority interests select an alternative. Further, participants reported that reason defaults enhance transparency, decision ease, and understanding of the choice relative to standard defaults and forced choices.

Keywords: defaults, choice architecture, heterogeneity, transparency, nudge

Statement of Relevance:

Defaults pre-select an option that is chosen automatically unless a decision-maker selects an alternative, and are often applied in consequential domains (e.g., organ donation, retirement savings). Defaults are effective at boosting the choice of the pre-selected option; however, they are often not designed to accommodate different options being better for different individuals. This is because defaults typically opt all decision-makers into the option that is better for the majority, potentially at the expense of individuals with minority interests. Compounding this issue, the mechanisms that make defaults effective (e.g., inertia, perceived recommendation) may keep people from opting out of a default when appropriate. In this paper, we investigate a default intervention designed to encourage those with minority interests to opt for an alternative, while guiding the majority to the default option. This intervention may allow defaults to function in consequential domains while reducing the costs imposed on individuals with minority interests.

INTRODUCTION

Choices with a default specify one option that is chosen automatically unless the decision-maker selects an alternative. For example, many employees are defaulted into participating in their company's 401k plan, so that they will accrue retirement savings even if they take no action. Defaults are powerful tools for nudging individuals towards a specific option and have been used to lift many ostensibly beneficial behaviors (Thaler & Benartzi, 2004; Thaler and Sunstein, 2008). For instance, countries that enroll citizens as organ donors by default tend to have far greater rates of organ donation (Johnson & Goldstein, 2003). Further, people are significantly more likely to get a flu vaccine when they are defaulted into a scheduled appointment instead of having to opt in (Chapman et al. 2010).

Choice architects typically select the option that is best for most people to be the default. For example, employees are defaulted into retirement savings programs because most people don't save enough for retirement (Beshears et al. 2016; Lusardi, 2001). However, people often have heterogeneous interests, meaning the default may not be the best option for all individuals (Sunstein, 2013). As a result, people with minority interests may be defaulted into suboptimal options (Beshears et al., 2016; Carroll et al., 2009; Choi et al., 2002; Cheryan & Markus, 2020). For example, in a study of 401K enrollment across three firms, 67.7% of employees thought they were saving too little, and only a third of these employees intended to save more in the near future. However, once a savings plan with a greater contribution was made the default option, uptake of the plan exceeded 85% (Choi et al., 2002), suggesting that this default led some individuals to save more than they intended. Other work has found that defaults often favor culturally masculine interests, potentially to the disadvantage of other genders (Cheryan & Markus, 2020).

Compounding this issue, individuals with minority interests may hesitate to opt out of a default option because they interpret it as a recommendation by an authority or an endowed option (among other possibilities; Johnson, 2022; Jachimowicz, Duncan, Weber, & Johnson, 2019; McKenzie et al., 2006). Further, those with lower income or less education may be more likely to stick with problematic defaults (Beshears et al., 2016; Sunstein, 2013). In one study, individuals were defaulted into an unusually high contribution rate. Although roughly 60% of individuals shifted from the default contribution rate to the lower (more manageable) rate, lower income individuals were more likely to stick with the default (Sunstein, 2013).

While past research has proposed interventions that could address people's heterogeneous interests, it is unclear if existing interventions will always be able to direct those with minority interests to better options. For example, research investigating the effect of making defaults more transparent has typically found that disclosing defaults does not make them less influential (Bruns et al. 2018; de Ridder, Kroese, & van Gestel, 2022; Loewenstein, Bryce, Hagmann, Rajpal, 2015; Michaelsen, Nyström, Luke, & Hedesström, 2021; Paunov, Wänke, & Vogel, 2019; Steffel, Williams, Pogacar, 2016). Such work has tested disclosures that inform decision makers of the intended effects of setting a default (to encourage choice of an option) and the default itself. Overall, this work suggests that increasing the transparency of the choice architecture may not be an effective way to guide those with minority interests to choose an alternative option as these disclosures don't reliably increase the selection of alternative options.

Another stream of research has proposed personalized defaults that are tailored to individuals' interests (Mills, 2022; Porat & Strahilevitz, 2014; Smith, Goldstein, & Johnson, 2013; Sunstein, 2013). Personalized defaults work by either using information about a decision-maker to predict which option is best for them, or collecting such information during the

decision-making process. However, personalized defaults are only a feasible solution if the choice architect can reliably predict which option is the best for each individual ex ante or collect predictive information about individuals' interests, which may not always be possible.

In this paper, we test whether presenting information about heterogeneity in interests alongside defaults can help individuals with minority interests select options that are better for them, and dub this intervention a "reason default." Reason defaults pre-select the option that is best for most individuals (like standard defaults), but they also present information that explains 1) why the default was selected and 2) when people should opt for an alternative option. The distinct goals of this information are to 1) communicate that there is heterogeneity in interests, 2) help people to understand when each alternative is an appropriate choice, and 3) reduce perceptions that only one option has special status, such as being recommended by an authority. Importantly, unlike personalized defaults, reason defaults present all participants with the same default and information. Thus, choice architects don't need to be able to predict individuals' interests to employ reason defaults.

We believe that reason defaults can maintain many of the advantages of standard defaults, while helping those with minority interests opt for an alternative. Because reason defaults still have a default option and communicate that the default is best for the majority of individuals, we believe that they are likely to lead more individuals to pick the default option than choices without a default. However, because reason defaults alert decision makers that an alternative option may be the better choice for them, individuals may be less likely to feel that they are disregarding a recommendation from an expert or missing out on a more favorable option if they opt out of the default. Further, because reason defaults increase transparency regarding when each option is beneficial (in contrast to past interventions that increase

transparency about the choice architecture itself), they may help individuals who are uninformed about the options, are uninformed about their interests, or are unable to infer which option aligns with their interests. We test these propositions in five preregistered experiments.

OPEN SCIENCE

For each study, we set the sample size before any data were obtained. We report all exclusions (if any), all manipulations, and all measures in the article or in the supplemental materials. We have posted preregistrations, raw data, analysis code, original materials, and supplemental materials for all studies on this project's OSF page: (https://osf.io/zjqxd/?view_only=d876fdccd2804e9eb61f91988f2490a4). We focus on the key analyses in the paper and present secondary analyses in supplemental materials S1 through S5.

METHODS

In Studies 1a-1c & 2, we used incentivized between subjects experiments to test whether reason defaults can help decision makers select options that are in line with their interests. We designed these studies to give participants the clear goal of maximizing their earnings from participating, and thus, define the option that maximizes each participants' earned incentive as the option most in line with their interests. Participants having transparent interests was necessary in order to investigate our research questions, but we should note that understanding which option best aligns with a decision makers' interests is not as straightforward in many real world domains (see the General Discussion for further discussion of this point). In Study 3, we explore people's beliefs about reason defaults, standard defaults, and forced choices in order to

understand how people perceive reason defaults relative to more common choice architecture tools.

Study 1a

In this study, participants completed a simulated savings task. Participants put the $1 base payment they earned for participating into a savings account for a simulated 20 year period. Their final study pay depended on the interest and tax accrued during the 20 year scenario. Participants were randomly assigned to one of two future tax changes (likely to increase or likely to decrease) and one of five default conditions (Forced choice (i.e., no default), Standard Traditional IRA default, Standard Roth IRA default, Reason Traditional IRA default, & Reason Roth IRA default). We predicted that those who received the reason default message would be most likely to choose a savings plan consistent with their most likely future tax rate.

Participants and procedures

Study 1a used a 5 (choice structure conditions: Forced choice, Standard default Roth, Standard default Traditional, Reason default Roth, Reason default Traditional) factor between subjects design. We preregistered that we would recruit 1000 participants (200 per condition) on Amazon Mechanical Turk. We only recruited participants from the United States. Individuals completed a captcha, gave consent, and completed an attention check to begin the study. Participants were paid $1.00 to participate and had the opportunity to earn a bonus. 1153 individuals clicked on the study link. 78 individuals failed the attention check and were not allowed to move on to the study. 171 responses were associated with a participant ID that appeared more than once, and we preregistered that responses with repeat IDs would be excluded

from our sample. 938 individuals with distinct participant IDs answered the main dependent variable. This sample was 62% male and 37 years old on average.

After passing the attention check, participants were prompted to read carefully as their decisions would determine their bonus. They also learned they could earn an extra bonus of $0.20 by answering two comprehension questions correctly at the end of the study. Then, participants read about the simulated savings task they would take part in.

Participants read that they would start with a balance of $1 and read about different savings plans into which they could invest their balance for a hypothetical 20 year period. The plan they chose would affect their earnings because they would need to pay taxes and their base pay would accrue interest over the 20 year period. At the end of the study, they would receive their base pay and any accrued interest after tax in real money as a bonus. We included a note explaining the simulative nature of the study to help tie participant tax expectations to the information in the survey: “Please note that this survey is an economic simulation. The interest and tax rates in this scenario do not reflect the real world rates, and different people will get assigned different rates. Please pay close attention to the instructions for information on your rates.”

Next, participants learned their base pay and rates in the scenario (including the interest rate, their current tax rate, and the probable change to their tax rate over the 20 year scenario). For all participants, the interest rate was fixed at 3% and the starting tax rate was fixed at 20%. Participants were randomly assigned to one of two tax change regimes, either reading that their tax rate was likely to increase (“Over the next 20 years, there is a 90% chance your tax rate will increase and a 10% chance it will decrease”) or likely to decrease (“Over the next 20 years, there is a 90% chance your tax rate will decrease and a 10% chance it will increase”). We assigned

participants heterogeneous circumstances (increasing versus decreasing tax rates) so that we could test how supportive reason defaults, standard defaults, and forced choices are of heterogeneous interests.

Next, participants were reminded of the simulation task and rules. Following the reminder, they read about two different types of US retirement savings plans, a Roth IRA (in which taxes are paid upfront and after tax dollars earn interest) and a Traditional IRA (in which pretax dollars are saved and taxes are taken out at the time of withdrawal), in random order. Descriptions of the plans were based on Vanguard's website detailing IRAs at the time of the study (see Table 1). Given the tax rules for these IRAs, participants who expected their tax rate

to decrease would earn more by choosing a Traditional IRA, and participants who expected their tax rate to increase would earn more by choosing a Roth IRA.

IRA type	Description
Roth	With a Roth IRA, you get a future bonus: Every penny you withdraw from your savings stays in your pocket, not Uncle Sam's. A Roth IRA is an individual savings account that offers tax-free growth and tax-free withdrawals. Roth IRA rules dictate that after the allotted period you can withdraw your money when you want to and you won't owe any taxes. You'll pay taxes on your contributions now, and you'll never pay taxes on withdrawals of your Roth IRA contributions and your earnings after the allotted period.
Traditional	Want to put off your tax bill while you save? Think traditional IRA. A traditional IRA is a type of individual savings account that lets your earnings grow tax-deferred. You pay taxes on your investment gains only when you make withdrawals after the allotted period. You won't pay taxes on your contributions now, and you'll pay ordinary income tax on withdrawals of all traditional IRA earnings.

Table 1. IRA Descriptions in Study 1.

After reading the descriptions, participants chose between the two plans. Participants were randomly assigned to see one of five default structures when choosing between the plans. In the forced choice group, participants were prompted to choose one of the two plans with no additional information. In the standard default Traditional IRA group, participants saw that the Traditional IRA is preselected and read a recommendation, “We recommend the Traditional IRA.” Similarly, in the standard default Roth IRA group, participants saw that the Roth IRA is preselected and read a recommendation, “We recommend the Roth IRA.” In the reason default Traditional IRA condition, participants saw that the Traditional IRA is the default option and

read, “We recommend the Traditional IRA because many people expect their taxes to be lower in the future. However, people who expect their tax rate to be higher in future are best served by the Roth IRA.” Those in the reason default Roth IRA condition saw that the Roth IRA is the default option and read, “We recommend the Roth IRA because many people expect their taxes to be higher in the future. However, people who expect their tax rate to be lower in future are best served by the Traditional IRA.”

After the choice task, participants reported whether they expected their tax rate to be higher before or after the imagined 20 year period, and completed three exploratory measures asking about their choice of savings plan (described at the end of Supplement 1). Next, participants answered two comprehension check questions that probed understanding of the study’s instructions regarding the tax regime (“Which statement accurately represents how your tax rate could change after the 20 year period in this scenario?”, “What is your tax rate at the beginning of the scenario?”) with four multiple-choice options each. 76% and 89% of participants answered each question correctly, respectively. Finally, participants learned their rates and balance at the end of the 20 year scenario (“Please imagine 20 years have passed. Now, the interest rate is 3%. Your current tax rate is [30, 10]%. Your tax rate has [increased, decreased]. Your current holdings are $[final balance].” Participants completed demographic questions (sex, age, and whether English is one’s native language) to complete the survey.

Results

We collapsed the two standard default conditions and the two reason default conditions leaving three conditions: forced choice, standard default, and reason default. Then, we ran chi-square tests comparing each possible pairing of these three conditions on the percent of choices participants made that are consistent with their most likely future tax rate excluding participants

who failed either comprehension check question (see Supplement 1 for the results with other specifications). Choosing the Roth IRA is consistent with a higher expected tax rate in the future since one would pay a lower tax rate (ending the study with higher earnings) by paying sooner. Choosing the Traditional IRA is consistent with a lower expected tax rate in the future since one would pay a lower tax rate by paying later.

We found that the reason default structure helped participants pick the IRA that was consistent with their most likely future tax rate. 74% of participants chose an IRA consistent with their assigned future tax rate (e.g., choosing a traditional IRA when taxes were likely to decrease) in the reason default condition compared to 62% of participants in the standard default condition, $\chi2(1, N = 556) = 10.22$, $p = .001$, $\phi_{cramer} = 0.14$, and 63% of participants in the forced choice condition, $\chi2(1, N = 408) = 5.38$, $p = .020$, $\phi_{cramer} = 0.11$. Reason defaults were especially helpful to participants who faced a "bad" default that was inconsistent with their incentives. Participants who were assigned to a "bad" default were significantly more likely to override this default in the reason default condition (68%) than the standard default condition (48%), $\chi2(1, N = 277) = 11.55$, $p = .001$, $\phi_{cramer} = 0.20$, while participants in the reason (80%) and standard (75%) default conditions were similarly likely to accept a "good" default, $\chi2(1, N = 279) = 1.34$, $p = 0.248$, $\phi_{cramer} = 0.07$. The results are consistent when analyzing the percentage of participants who chose the plan consistent with their beliefs instead of their most likely future tax rate, and including participants who did not pass both comprehension check questions (see Supplement 1).

Studies 1b & 1c

Studies 1b and 1c are follow-ups to Study 1a that assign more participants to subsets of the conditions in Study 1a and a new condition, which provided the reason message without a

default, allowing us to separate the effect of supplying information about heterogeneity from defaults. All participants who received a default option in both studies were assigned to a Roth IRA default, participants in Study 1b were assigned to a 90% chance of their taxes decreasing (making the Roth IRA a "bad default"), and participants in Study 1c were assigned to a 90% chance of their taxes increasing (making the Roth IRA a "good default"). These designs allowed Studies 1b & 1c to replicate the results of Study 1a, and test the effects of reason defaults when the default option is "good" versus "bad" with more statistical power. Further, because most participants in Study 1a expected taxes to increase (57%) and chose the Roth IRA (65%), these specific conditions allowed us to test the effect of reason versus standard defaults under potentially the best and worst circumstances for standard defaults: when the default option is the intuitive but inferior option (and reason defaults may have the most to offer) in Study 1b, and when the default option is the intuitive and advantageous option (and reason defaults may have the least to offer) in Study 1c.

Participants and procedures

Studies 1b & 1c both used a 2 (default: present, not present) X 2 (reason message: present, not present) between subjects design, creating four conditions: Forced choice, Standard default Roth, Reason default Roth, Reason message (with no default). We preregistered that we would recruit 800 participants (200 per condition) on Amazon Mechanical Turk for both studies, and both used CloudResearch approved participants. We only recruited participants from the United States. Individuals completed a captcha, gave consent, and completed an attention check to begin the study. Participants were paid $1.00 to participate and had the opportunity to earn a bonus. 867 & 863 individuals clicked on the study link in studies 1b & 1c respectively. 20 & 32 individuals failed the attention check and were not allowed to move on to the study. 76 & 42

responses were associated with a participant ID that appeared more than once, and we preregistered that responses with repeat IDs would be excluded from our sample. 766 & 783 individuals with distinct participant IDs answered the main dependent variable. These samples were 55% & 47% male and 40 & 42 years old on average.

The procedures of Studies 1b & 1c were the same as Study 1a with the exception of the following changes. First, all participants in Study 1b were assigned to a tax regime in which their taxes had a 90% chance of decreasing and a 10% chance of increasing, and all participants in Study 1c were assigned to a tax regime in which their taxes had a 90% chance of increasing and a 10% chance of decreasing. This was the only difference between Study 1b & 1c.

Second, in the conditions exposed to a default option (Standard default & Reason Default), all participants were assigned to a Roth IRA default. Third, these studies included a new "Reason message" condition which included the information from the Reason default condition without the default. Participants assigned to the Reason message condition read one sentence describing when Roth IRAs are advantageous and one sentence describing when Traditional IRAs are advantageous in counterbalanced order (e.g., "People who expect their tax rate to be higher in the future are best served by the Roth IRA. People who expect their tax rate to be lower in the future are best served by the Traditional IRA.").

Finally, we did not collect participants' beliefs about their future tax rate after they chose a savings plan or the three exploratory measures asking about participants' choice of plans. Thus,

the only dependent variable in these studies was choice consistent with one's most likely future tax rate.

Results of Study 1b – Bad Default

We preregistered that we would run a logistic regression of choice of the Traditional IRA (the better option given the tax regime) on a contrast coded (-1, 1) "default" term (indicating whether a default was present), a contrast coded (-1, 1) "reason message" term (indicating whether participants were presented with the information about their options on the choice screen), and the interaction between these two terms. The results presented below only include participants who answered both comprehension check questions correctly. The results are consistent when including all participants (see Supplement 2).

The logistic regression revealed that participants assigned to a bad default selected the savings plan consistent with their tax regime (i.e., the Traditional IRA) significantly less often (46%) than those assigned to no default at all (55%), $ß$ = -0.20 (95% CI: -0.37, -0.04), z = -2.41 p = .016. This result is consistent with the notion that bad defaults, which nudge people towards options that are inconsistent with their best interests, can negatively affect decision-makers. In contrast, participants assigned to receive the reason message selected the advantageous Traditional IRA significantly more often (63%) than those assigned not to receive this message (38%), $ß$ = 0.51 (95% CI: 0.35, 0.68), z = 6.05, p < .001. This result is consistent with the notion that the reason message helped participants choose an option that was consistent with their interests. These two main effects were not qualified by a significant interaction, $ß$ = -0.09 (95% CI: -0.27, 0.065), z = -1.19, p = .236.

Turning to the individual conditions, participants who received the bad default were significantly more likely to select the advantageous savings plan when they received the reason

message (56%) than when they did not (36%), $\chi 2(1, N = 305) = 12.38$, $p < .001$, $\phi_{cramer} = 0.20$. Thus, consistent with the results of Study 1a, we found strong support for the notion that reason defaults help people to make better choices when they are assigned to a bad default that does not reflect their interests. Next, we test the effects of defaults and reason messages when defaults do reflect people's interests.

Results of Study 1c – Good Default

We preregistered that we would run a logistic regression of choice of the Roth IRA (the better option given the tax regime) on a contrast coded (-1, 1) "default" term, a contrast coded (-1, 1) "reason message" term, and the interaction between these two terms similar to Study 1b. The results presented below only include participants who answered both comprehension check questions correctly. The results are consistent when including all participants (see Supplement 3).

The logistic regression revealed that participants assigned to a good default selected the savings plan consistent with their tax regime (i.e., the Roth IRA) significantly more often (93%) than those assigned to no default at all (88%), $ß = 0.28$ (95% CI: 0.01, 0.55), $z = 2.03$, $p = .042$. This result supports the notion that properly selected defaults can improve decision makers' outcomes, consistent with the large body of research demonstrating the effectiveness of defaults as a choice architecture tool. In contrast, participants who received the reason message selected the Roth IRA at a similar rate (91%) as those who did not receive this message (90%), $ß = 0.01$ (95% CI: -0.26, 0.27), $z = 0.05$, $p = .961$. Thus, under the conditions in Study 1c, the reason

message did not significantly improve participants' outcomes. These two main effects were not qualified by a significant interaction, *ß* = -0.12 (95% CI: -0.39, 0.14), *z* = -0.91, *p* = .365.

Turning to the individual conditions, we once again replicated past evidence that defaults can be a beneficial choice architecture tool, as participants in the standard default condition chose the advantageous Roth IRA more often (94%) than participants in the forced choice condition (87%), $\chi 2(1, N = 345) = 4.47$, $p = .034$, $\phi_{cramer} = 0.11$. Importantly, although we did not find evidence that the reason message was an effective intervention on its own, we also did not find evidence that the presence of the reason message decreased the effectiveness of the default; as participants in the standard default (94%) and reason default (92%) conditions selected the more advantageous Roth IRA at similar rates, $\chi 2(1, N = 334) = 0.30$, $p = .585$, $\phi_{cramer} = 0.03$. Thus, consistent with the results of Study 1a, we found evidence that standard and reason defaults may be similarly effective in leading people to accept a "good" default.

Discussion

Collectively, the results of Studies 1a, 1b, and 1c demonstrate that reason defaults may be a particularly effective intervention for guiding decision makers with heterogeneous interests towards appropriate options while maintaining the benefits of standard defaults. Studies 1a and 1b found that reason messages are beneficial alongside defaults when those defaults happen to pre-select the inferior option for a particular individual. Indeed, participants were 56% more likely to select the better option given their incentives when the bad default in Study 1b was accompanied by a reason message. However, defaults typically select the better option for most individuals, and the results Study 1c suggest that informational messages alone without a default may forgo the long established benefits of properly selected defaults. Finally, Studies 1a and 1c found that reason defaults and standard defaults were similarly effective in leading participants

to select a good default, which suggests that augmenting a standard default to include a reason message may not diminish its effectiveness.

Study 2

In Study 2, we test whether reason defaults can help participants to make better choices when there is natural heterogeneity in their interests, as opposed interests assigned by experimenters. Participants were randomly assigned to see one of three choice structures (forced choice, standard default, reason default) when choosing between two incentivized 10-question quizzes for a bonus. We used pretests to choose a default test that is better for most people, and an alternative test that is better for a minority of the population. We hypothesized that those who saw information about when to choose each option via the reason default would score higher on the quiz they chose than those assigned to a standard default or a forced choice. We also hypothesized that individuals exposed to the reason message would be more likely to feel they had chosen the best option for themselves than those in the other conditions.

Participants and procedures

Study 2 used a 3 (choice structure conditions: forced choice, standard default, reason default) factor between subjects design. We preregistered that we would recruit 900 participants (300 per condition) on Amazon Mechanical Turk for Study 2. Individuals completed a captcha, gave consent, and completed an attention check to begin the study. Participants were paid $0.40 to participate and had the opportunity to earn a bonus. 1151 individuals clicked on the study link. 194 individuals failed the attention check and were not allowed to move on to the study. 176 responses were attached to a participant ID that appeared more than once, and we preregistered that responses with repeat IDs would be excluded from our sample. 834 individuals with distinct

participant IDs answered the main dependent variables. This sample was 48% male and 38 years old on average.

After passing the attention check, participants read about the timed quiz task and the bonus structure. Participants read: “Thank you for participating in this survey. Next, you will choose a test to take as part of the survey. Please try to answer as many questions as you can correctly. You will have 60 seconds to answer the questions. You will receive a $0.02 bonus for each question you answer correctly. Please don’t look up the answers!”

Next, participants were asked to choose one of two tests, labeled Test A and Test B, to take as part of the study. Test labels were counterbalanced, as was the order of the tests. The subjects of the tests were obscured because we designed this task to be an extreme case reflecting real world situations in which decision makers are unsure which option best matches their interests and choice architecture has the potential to be most helpful (e.g., choosing among retirement savings options, medical treatments, insurance plans).

We ran pretests to choose the subjects of these obscured tests. The default test was an English vocabulary and grammar test on which most people score better than chance but far from perfectly because of some difficult questions. The alternative test was a test on the TV show “Parks and Recreation.” Pretests showed that a small percentage of the population reported that they knew trivia about this topic very well, while the majority of the population reported not knowing trivia about this topic at all. This configuration of tests created a default which was best for the majority of the population (the English vocabulary and grammar test), and an alternative that was the better option for the minority of the population (the Parks and Recreation trivia test). This situation mimics the real life scenarios in which we believe reason defaults will be most

useful; situations in which the default has been selected to represent the interests of the majority, while a minority of the population is best served by the alternative.

Participants were randomly assigned to one of three default conditions. Those randomly assigned to the forced choice condition chose between Test A and Test B with no additional information about the options, reading "Please choose the test you would like to take as part of the survey." Those assigned to the standard default condition saw that one of the two options, always corresponding to the English vocabulary and grammar test, was pre-selected, and read: "We recommend the preselected option, as most people are better at answering the questions in [the default test]." Those in the reason default group also saw that the default option was pre-selected and read: "We recommend the preselected option, as most people are better at answering questions in [the default test]. However, people who are very familiar with the show Parks and Recreation do better on [the alternate test]." Thus, the reason default disclosed the key piece of information describing exactly who should opt out of the default option.

After making a choice, participants were reminded of the bonus for each correct answer and were informed that moving to the next screen would start the quiz and the 60 second timer. We implemented the 60 second timer so that looking up the answers to many or all questions would not be feasible. Once the quiz started, a timer at the top of the page began to count down from 60 seconds. Participants answered as many of the 10 available questions as they desired before the page automatically advanced after 60 seconds. Participants were able to advance the page themselves before the 60 second timer expired as well.

After taking the test and before learning their bonus, participants were reminded of the test they selected and informed of the topic of each test. Next, participants were asked whether they thought the test they took was the best option for them (choosing between yes or no) and

how confident they were that they could answer questions on the topic of each of the possible tests correctly (on a three point scale from 1, not at all confident, to 3, very confident). Participants reported demographic information (sex, age, and whether English is one's native language) to complete the study.

Results

We found that the reason default improved participants' outcomes. Participants assigned to a reason default answered more questions correctly (M=5.17, SE=0.14), earning larger bonuses, than participants assigned to a forced choice (M=4.35, SE=0.15), $t(549) = 3.93$, $p < .001$, $d = 0.33$, or standard default (M=4.45, SE=0.11), $t(558) = 4.01$, $p <. 001$, $d = 0.34$. Participants who faced a forced choice and standard default scored similarly, $t(555) = 0.50$, $p = .619$, $d = 0.04$. The reason default was effective because it helped the right participants select into the alternative test. Among participants who chose the alternative test (53% in the forced choice condition, 15% in the standard default condition, and 20% in the reason default condition, $\chi 2(2, N = 834) = 116.95$, $p < .001$, $\phi_{cramer} = 0.37$), participants assigned to a reason default scored higher (M=7.62, SE=0.43) than those assigned to a forced choice (M=4.03, SE=0.25), $t(199) = 7.38$, $p < .001$, $d = 1.17$, or standard default (M=3.71, SE=0.42) , $t(95) = 6.42$, $p < .001$, $d = 1.31$ (see Figure S2). Participants who chose the default test performed similarly across all conditions, $M_{forced}=4.72$, $SE_{forced}=0.16$; $M_{standard}=4.58$, $SE_{standard;}=0.11$; $M_{reason}=4.57$, $SE_{reason}=0.11$; $F(2, 588) = 0.38$, $p = .684$, $\eta^2_p = 0.001$.

Cumulative distributions of participants' scores across conditions give a more nuanced view of the tradeoff between the standard default and forced choice (see Figure 1). The standard default led the vast majority of participants (85%) to choose the default English vocabulary and grammar test, on which most participants perform moderately. The forced choice led roughly half

(53%) of participants to choose the alternative test, on which many participants perform poorly but some participants perform very well. As a result, a choice architect choosing between a forced choice and a standard default faces a tradeoff: The standard default guides everyone to the same lower variance alternative that is better for the majority of the population, while the forced choice results in more participants taking a higher variance alternative with greater upside for a minority of the population. This tradeoff is represented in Figure 1 by the grey (standard default) and black (forced choice) CDFs crossing around a proportion of .75. That is, the standard default led to better performance for roughly 75% of the population and the forced choice led to better performance for roughly 25% of the population.

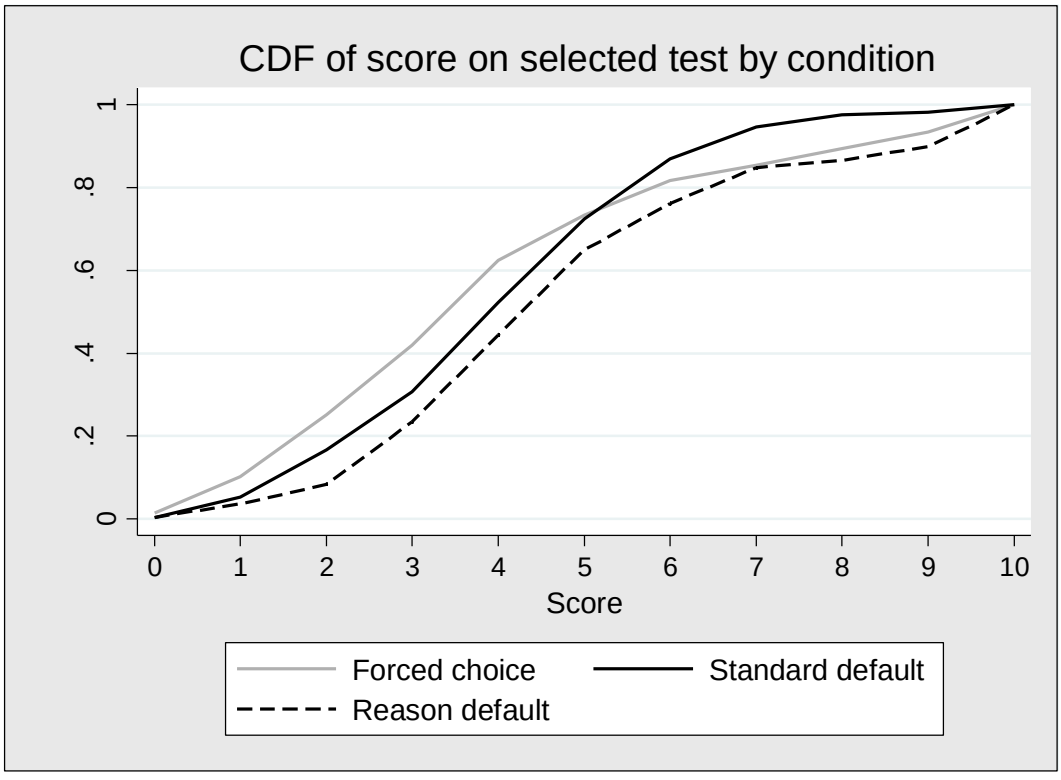


Fig. 1. Participants' scores on the test in Study 2. Participants completed a 10-item test as the main dependent measure of Study 2. This figure shows CDFs of the scores that participants got on their selected test across conditions. Points to the right represent higher scores and better performance.

The reason default condition demonstrates that choice architects don't need to make the aforementioned tradeoff. The reason default led the majority of participants (80%) to choose the

default test that is better for most, while leading the right minority of participants (20%) to opt into the alternative test. In other words, this intervention offered the best of both worlds, represented in Figure 1 by the dashed black CDF dominating the grey and black CDFs (i.e., being shifted completely to the right). This means that the reason default led to better performance for the entire population of participants, and that there is no subset of the population that would be better served by a forced choice or standard default in expectation (while the same cannot be said for either the forced choice or standard default).

After completing their chosen test and learning the subject of both tests, participants reported whether they believed that the test they selected was the best option for them. More participants thought they had taken the best test for them in the reason default condition (81%), than those in the standard default condition (53%, $\chi 2(1, N=560) = 48.70$, $p<0.001$, $\phi_{cramer} = 0.29$) and forced choice condition (47%, $\chi 2(1, N=551) = 66.70$, $p<0.001$, $\phi_{cramer} = 0.35$). There was no significant difference in thinking one had taken the best test between the forced choice and standard default conditions ($\chi 2(1, N=557) =1.73$, $p=0.188$, $\phi_{cramer} = 0.06$). Overall, the results of Study 2 suggest that a reason default message can lead more people to select a default option that is better for most people than a forced choice, while simultaneously helping those with minority interests opt for an alternative option.

Study 3

In Study 3, we explore people's perceptions of reason defaults, standard defaults, and forced choices to better understand how laypeople feel about these different interventions. In this study, participants read about a group of policy makers planning to present a choice to their constituents in one of six contexts (diet, environmental sustainability, savings, airbags, privacy, and organ donation). Participants saw an example of each of three choice structures (a forced

choice, a standard default, or a reason default) applied to their assigned context in random order. After reading about each choice structure, participants answered 14 questions corresponding to seven different topics (e.g., decision ease, strength of recommendation, decision maker autonomy, transparency). This allowed us to test how people's perceptions of reason defaults compare to their perceptions of standard defaults and forced choices.

Participants and procedures

Study 3 used a 3 (choice structure, within) x 6 (context, between) mixed design. We preregistered that we would recruit 900 participants on Amazon Mechanical Turk (150 per context). We randomized the order in which participants saw three choice structures within their assigned context. Individuals completed a captcha, gave consent, and completed an attention check to begin the study. Participants were paid $0.85 to participate. 1151 individuals clicked on the study link. 50 individuals failed the attention check and were not allowed to move on to the study. 65 responses were attached to a participant ID that appeared more than once. 916 participants answered the main dependent variables. We preregistered that responses with duplicated IDs would be excluded from our sample. 889 individuals with distinct participant IDs answered the main dependent variables. This sample was 54% male and 41 years old on average.

First, participants learned that they would read about three different ways a policy maker could frame a decision. Participants were assigned to one of six decision contexts: Savings, where the choice was between a Roth IRA or traditional IRA; Organ donation, where the choice was between becoming an organ donor or not becoming an organ donor; Diet, where the choice was between a side of fries or apple slices; Airbags, where the choice was whether to add an

airbag on-off switch to a car; Sustainability, where the choice was between CFL or LED light bulbs; and Privacy, where the choice was between allowing or disallowing internet cookies.

After reading about the task, participants read about each option that would be included in the choice structures as well as the consequences of the decision. For example, a participant assigned to the Diet context would read: “Fries and apple slices are two different side dish options that have different health implications. Specifically, with fries, you choose a food that is suggested to be an occasional treat for most people, and take in high quantities of fat, sodium, and overall calories. With apple slices, you choose a food that is suggested to be consumed regularly for most people, and take in low quantities of fat, sodium, and overall calories. This is a consequential choice because it can impact one's overall health as well as one's satisfaction with food.”

After reading about the options, participants were reminded they would see multiple choice structures and informed they would provide ratings for each one: “Next, you will see 3 different ways that the restaurant can frame this decision for customers, and complete multiple ratings of each one.” Then, they viewed each of three choice structures, a forced choice, a standard default, and a reason default, in random order.

For each structure, participants were prompted to imagine a group of choice architects deciding how to structure a choice. For example, those assigned to the Diet context read: “Imagine a small group of managers at a restaurant who make decisions about food. One decision this group has to make regards how to handle situations in which people are deciding

between side dish options at fast food restaurants. They are considering setting up a choice in the following way." The choice structure was visible below this prompt.

For the reason default and standard default, the survey randomly determined which option was the pre-selected default and what order the options were presented in for each participant. Participants assessing a reason default read, "We recommend [Default] because many people [Default reason]. [Alternative explanation]" (see Table 2 for the phrasing for each condition). When assessing a standard default, participants read, "We recommend the [Default]."

When reading about the forced choice structure, participants viewed a choice between two options without a pre-selected option or any recommendation.

Domain	Default	Default reason	Alternative explanation
Diet	Fries	like to eat tasty foods when dining out or need to gain weight	However, people who like to eat lower calorie foods for health reasons or prefer how apples taste are best served by the apple slices
	Apple slices	like to eat lower calorie foods for health reasons or prefer how apples taste	However, people who like to eat tasty foods when dining out or need to gain weight are best served by the fries
Sustainability	CFL bulbs	want to pay less for light up front	However, people who want to save on energy costs or lower their energy use in the long term are best served by LEDs
	LED bulbs	want to save on energy costs or lower their energy use in the long term	However, people who want to pay less for light up front are best served by CFLs
Savings	Roth IRA	expect their taxes to be higher in the future	However, people who expect their tax rate to be lower in the future are best served by the Traditional IRA
	Traditional IRA	expect their taxes to be lower in the future	However, people who expect their tax rate to be higher in the future are best served by the Roth IRA
Airbag	No on-off switch	who sit in the front seats of a car are of moderate height	However, those who expect people shorter than 4'11" or taller than 6'3" to sit in the front seats are best served by including an on-off switch
	Include on-off switch	who sit in the front seats of a car are shorter than 4'11" or taller than 6'3"	However, those who do not expect people shorter than 4'11" or taller than 6'3" to sit in the front seats are best served by not including an on-off switch as it ensures that the airbags will be active
Privacy	Accept cookies	want content tailored to them as they browse the site	However, people who want to limit the amount of information they share are best served by not accepting cookies
	Do not accept cookies	want to limit the amount of information they share	However, people who want content tailored to them as they browse the site are best served by accepting cookies
Organ donation	Donate	trust medical institutions and want to donate their organs	However, people who distrust medical institutions are best served by not becoming organ donors
	Do not donate	distrust medical institutions or do not want to donate their organs	However, people who trust medical institutions are best served by becoming organ donors

Table 2. This table presents the stimuli used for Study 3. The reasons listed for choosing each option describe the stimuli we chose for this study, but there are many more reasons why a decision maker may prefer each of the listed options. For example, there are many potential reasons why people may not want to be organ donors.

After viewing each choice structure, participants answered 14 questions – two questions on each of seven topics, on 5-point scales from "Strongly disagree" to "Strongly agree." The 7 topics were: to what extent the structure promotes understanding of the reasons for choosing each option, how effortful the choice is, whether the structure gives a strong recommendation, whether choosing one option is the norm, whether choosing the alternative is encouraged, to what extent a decision maker would feel autonomy, and whether the choice architect's intentions are transparent. All questions were presented in randomized order on the same page. See Table 3 for the exact wording of all questions.

Construct	Questions	
Promotes understanding	"It is clear why someone should choose each of the options"	
	"It is clear how the options match up with different preferences"	
Makes decisions easier	"Decision makers will have an easier time making a decision when the options are presented this way"	
	"The way the options are presented reduces the effort needed to make a good decision"	
Gives strong recommendation	"The way the decision is presented makes a strong recommendation"	
	"The people who set up this decision favor one option over another"	
One option is the norm	"One of the options is the status quo or default"	
	"Choosing one of the options is against the norm"	
Choosing the alternative is encouraged	"The way the choice is presented will lead everyone to make the same decision"	Reverse coded
	"The way the choice is presented will help people who have a minority preference (i.e. a preference opposite of the majority) to make the better choice for themselves"	
Decision makers have autonomy	"Decision makers will feel like they are making this decision for themselves"	
	"This decision is being made for decision makers to some extent"	Reverse coded
Choice architect intention transparent	"It is clear why the people who set up the decision presented the options in the way they did"	
	"The thinking of the people who set up the decision is transparent"	

Table 3. Question wording for Study 3.

Participants completed the process of reading about a choice setup and answering the 14 questions described above three times – one for each choice structure (forced choice, standard default, and reason default) in randomized order. After completing this process for all three choice structures, participants reported demographics (sex, age, and whether English is one's native language) to complete the survey.

Results

As preregistered, we collapsed all domains, defaults, and option orderings down to the three choice structures, leaving 3 conditions: reason default, standard default, and forced choice. Further, we averaged the two questions on each of the seven topics we investigated to create seven dependent measures (see Table S1). To analyze these measures, we ran paired t-tests comparing each pairing of the three conditions for each of the 7 aggregate measures. All means are significantly different from each other at $p<=.015$ & $d>0.08$ using paired t-tests. All of these differences remain statistically significant after applying Holm–Bonferroni correction for multiple comparisons as preregistered (see Table S2).

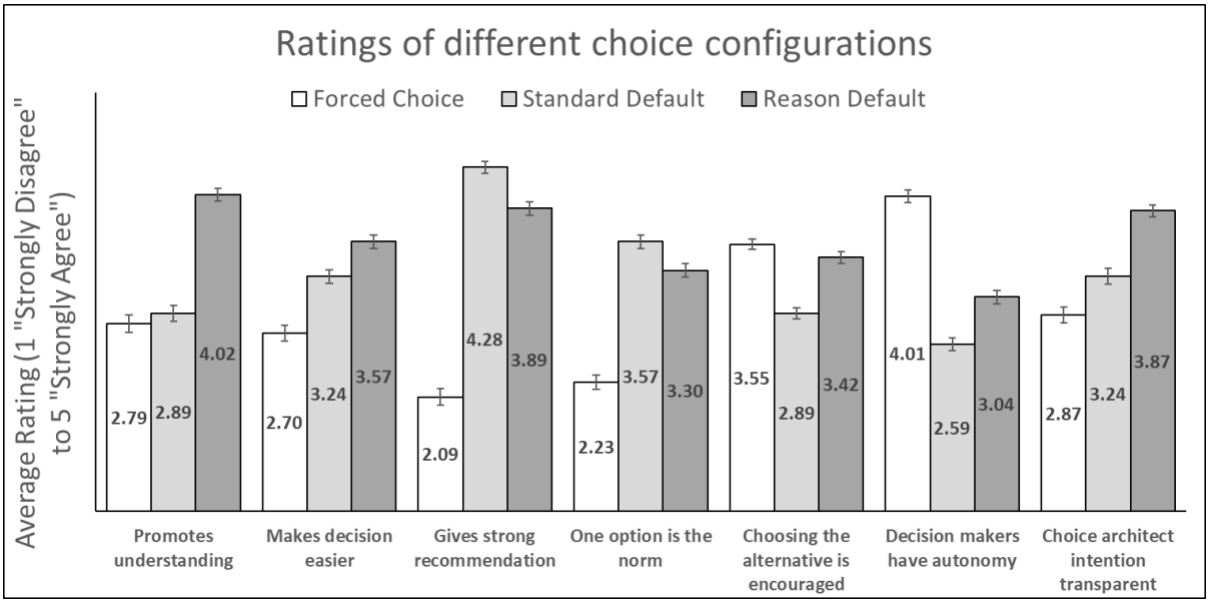


Fig. 2. Results for Study 3. Participants rated three different choice configurations in one of six domains on seven two-item scales. Error bars represent 95% CIs.

Participants reported that reason defaults have some advantages over both forced choices and standard defaults. Specifically, participants reported that reason defaults promote greater understanding of the options than standard defaults, $t(879) = 24.31$, $p < .001$, $d = 1.10$, and forced choices, $t(883) = 23.83$, $p < .001$, $d = 1.16$, (see Figure 2 for all means). Additionally, participants reported that reason defaults give greater transparency regarding the thinking of the choice architect than standard defaults, $t(879) = 14.01$, $p < .001$, $d = 0.60$, and forced choices, $t(883) = 20.50$, $p < .001$, $d = 0.98$. These results suggest that reason defaults may give people more insight about their options and the structure of their choice than both standard defaults and forced choices. Additionally, participants reported that reason defaults made decisions easier than both standard defaults, $t(879) = 8.58$, $p < .001$, $d = 0.33$, and forced choices, $t(883) = 17.01$, $p < .001$, $d = 0.82$. Overall, these findings suggest that relative to standard defaults and forced choices, reason defaults may be able to better inform people about their options without making their choices more difficult or confusing.

Participants also reported that reason defaults fell between standard defaults and forced choices on the other four dimensions that they rated. Specifically, participants reported that reason defaults: give weaker recommendations than standard defaults, $t(879) = -10.12$, $p < .001$, $d = 0.43$, but stronger recommendations than forced choices, $t(883) = 34.57$, $p < .001$, $d = 1.68$; leave decision makers more autonomy than standard defaults, $t(879) = 11.92$, $p < .001$, $d = 0.49$, but less autonomy than forced choices, $t(883) = -22.38$, $p < .001$, $d = 1.02$; indicate that one option is the norm less strongly than standard defaults, $t(879) = -8.48$, $p < .001$, $d = 0.28$, but more strongly than forced choices, $t(883) = 23.96$, $p < .001$, $d = 1.05$; and encourage choice of the alternative more strongly than standard defaults, $t(879) = 14.85$, $p < .001$, $d = 0.63$, but less strongly than forced choices, $t(883) = -3.53$, $p < .001$, $d = 0.16$. We interpret these results as

suggesting that reason defaults may be better than forced choices at guiding individuals toward the option that is better for most, while not suggesting as strongly as standard defaults that all individuals should select the default option.

General Discussion

The results of five studies suggest that presenting a message describing information about heterogeneity alongside defaults can improve both decision outcomes and decision makers' experiences. In Studies 1a, 1b, 1c, & 2, we found that reason defaults (compared to forced choices and standard defaults) improved participants' earnings by helping them choose the option that was aligned with their incentives. Reason defaults were especially beneficial when participants were assigned to a "bad" default that did not reflect their interests, and similarly as effective as standard defaults when participants were assigned to a "good" default that did reflect their interests. The reason defaults in Studies 1a-c increased participants earnings even though participants in all conditions were supplied with the information needed to understand which option was aligned with their interests. We believe that messaging about heterogeneity at the time of choice may be more effective than simply providing educational information because it directly addresses the forces that can give defaults power, such as beliefs that only the default option is being strongly recommended by an authority.

In Study 3, participants expressed their beliefs about reason defaults, standard defaults, and forced choices. Participants reported that reason defaults encourage the choice of alternative options more than standard defaults, while still promoting the default option more than forced choices, and enhancing transparency, decision ease, and understanding relative to both forced choices and standard defaults. In line with the results of Studies 1a-c & 2, this suggests that reason defaults may have the potential to strongly encourage choice of the default option for

most individuals, while simultaneously encouraging those with minority interests to choose an alternative option. Further, the notion that reason defaults may increase understanding and decision ease while blunting the default recommendation may increase decision-makers' competence in the decision domain and likelihood of deliberating, similar to a "system 2 nudge" (see Sunstein, 2016), or "boost" (see Hertwig & Grüne-Yanoff, 2017).

Limitations & Future Directions

The empirical approach that we employed in this article does have limitations. One salient example is that we designed Studies 1 & 2 so that participants' interests were transparent, but it is often unclear what is "best" for decision-makers' in real-world domains. For instance, even experts sometimes disagree on issues such as dietary and health interventions. Furthermore, individuals may occasionally prefer options that experts deem detrimental (e.g., smoking). Real-world decisions also often involve weighing multiple interests simultaneously. A single food choice, for example, can impact health (e.g., low vs. high calorie), finances (e.g., expensive vs. cheap), environmental goals (e.g., locally grown vs. shipped), and moral values (e.g., vegan vs. animal products) among other considerations. Consequently, determining the "best" option may require not only understanding a decision-maker's many interests, but also understanding how they prioritize and balance diverse interests, which we believe is often an intractable problem.

However, the notion that we may often not understand what option is best for a decision-maker makes us *more* optimistic about the potential benefits of reason defaults in the real world. The results of Studies 1 & 2 suggest that reason defaults can help decision makers to opt out of defaults that don't align with their interests, which would be beneficial when choice architects misunderstand an individual's interests and pre-select an inferior option. Further, the results of Study 3 suggest that reason defaults may increase decision-makers' perceived autonomy relative

to standard defaults and increase their understanding of the options relative to both forced choices and standard defaults. As a result, reason defaults could encourage decision-makers to take a more active role in their choices, and consider which option is best when other parties don't or can't fully understand their interests. We look forward to future work that investigates these possibilities.

We also believe that reason defaults are likely to have limitations as an intervention. For example, reason defaults are most likely to be beneficial when people's interests exhibit heterogeneity, as a standard default may suffice when one option is best for everyone. Further, Study 3 suggests that standard defaults are seen as recommending the default option more strongly, which could be beneficial when there is one dominant option. Future research should test whether individuals are more likely to inappropriately deviate from reason defaults than standard defaults in some domains as a result, although we did not find evidence of this behavior in Studies 1a, 1c, and 2. Relatedly, standard defaults may also be a better option when inertia is a helpful (instead of harmful) feature of the choice architecture, as reason defaults may increase attention to the presence of the decision and the decision maker's likelihood of deliberating.

When comparing reason defaults to other choice architecture tools that could address heterogeneity, like personalization, choice architects likely still face tradeoffs. By providing an option that is predicted by previously collected information about an individual, personalization offers decision makers time savings, as they can forgo reading additional information or investigating their interests. However, time savings from using personalization come at the expense of decision makers gaining a better understanding of their options and feeling autonomy from paternalism. Thus personalization could be preferred to reason defaults in settings where decision makers do not wish to make decisions on their own or are not interested in the reasoning

behind an option. Such settings could involve decisions where a decision maker knows information about his interests and does not need to consider them (e.g. making a decision where options depend on one's measurements or making a decision using a set amount of money.) On the other hand, reason defaults may be better for settings where interests are more volatile over time and personalized predictions may be less than sufficiently accurate to one's present interests. (Reason defaults may also offer an option for individuals who prefer not to share personal data as predictions may be less reliable.).

Issues may also arise with reason defaults in cases where decision makers have biased beliefs. If decision makers have biased beliefs about what is beneficial, reason defaults could provide an opportunity for them to make an incorrect choice, one counter to their interests. For example, if one mistakenly believes that brighter lights are less efficient or makes an error in calculating energy costs, he may use the bulb choice reason message to select the more costly option. Biased beliefs could also detract from feelings that reason defaults offer transparency if decision makers feel that the reason message aims to manipulate their beliefs (e.g. one may believe that sugar is particularly bad for health and feel that listing it as the lower calorie option is misleading).

Separately, reason defaults could be hard to implement in domains in which there are many options because the reason default messaging may become too long or difficult to understand. We look forward to future work that explores the many potential benefits and limitations of reason defaults.

References

Beshears, J., Choi, J. J., Laibson, D., Madrian, B. C., & Wang, S. Y. (2016). Who is easier to nudge?. NBER Working Paper, 401. http://www.nber.org/papers/w23679

Bruns, H., Kantorowicz-Reznichenko, E., Klement, K., Jonsson, M. L., & Rahali, B. (2018). Can nudges be transparent and yet effective?. *Journal of Economic Psychology*, *65*, 41-59. https://doi.org/10.1016/j.joep.2018.02.002

Carroll, G. D., Choi, J. J., Laibson, D., Madrian, B. C., & Metrick, A. (2009). Optimal defaults and active decisions. *The quarterly journal of economics*, *124*(4), 1639-1674.

Chapman, G. B., Li, M., Colby, H., & Yoon, H. (2010). Opting in vs opting out of influenza vaccination. *Jama*, *304*(1), 43-44. https://doi.org/10.1001/jama.2010.892

Cheryan, S., & Markus, H. R. (2020). Masculine defaults: Identifying and mitigating hidden cultural biases. *Psychological Review*, *127*(6), 1022. https://doi.org/10.1037/rev0000209

Choi, J. J., Laibson, D., Madrian, B. C., & Metrick, A. (2002). Defined contribution pensions: Plan rules, participant choices, and the path of least resistance. *Tax policy and the economy*, *16*, 67-113. https://doi.org/10.1086/654750

de Ridder, D., Kroese, F., & van Gestel, L. (2022). Nudgeability: Mapping conditions of susceptibility to nudge influence. *Perspectives on Psychological Science*, 17(2), 346-359.

Hertwig, R., & Grüne-Yanoff, T. (2017). Nudging and boosting: Steering or empowering good decisions. *Perspectives on Psychological Science*, *12*(6), 973-986.

Jachimowicz, J. M., Duncan, S., Weber, E. U., & Johnson, E. J. (2019). When and why defaults influence decisions: A meta-analysis of default effects. *Behavioural Public Policy*, *3*(2), 159-186. https://doi.org/10.1017/bpp.2018.43

Johnson, E. J. (2022). *The elements of choice: Why the way we decide matters*. Penguin.

Johnson, E. J., & Goldstein, D. (2003). Do defaults save lives?. *Science, 302*(5649), 1338-1339. https://doi.org/10.1126/science.1091721

Loewenstein, G., Bryce, C., Hagmann, D., & Rajpal, S. (2015). Warning: You are about to be nudged. *Behavioral Science & Policy*, *1*(1), 35-42. https://doi.org/10.1353/bsp.2015.0000

Lusardi, A. (2001). Explaining why so many people do not save. Center for Retirement Research Working Paper, (2001-05). https://dx.doi.org/10.2139/ssrn.285978

McKenzie, C. R., Liersch, M. J., & Finkelstein, S. R. (2006). Recommendations implicit in policy defaults. *Psychological Science, 17*(5), 414-420. https://doi.org/10.1111/j.1467-9280.2006.01721.x

Michaelsen, P., Nyström, L., Luke, T. J., & Hedesström, M. (2021). Downstream consequences of disclosing defaults: influences on perceptions of choice architects and subsequent behavior. *Comprehensive Results in Social Psychology*, 1-24.

Mills, S. (2022). Personalized nudging. *Behavioural Public Policy, 6*(1), 150-159.

Paunov, Y., Wänke, M., & Vogel, T. (2019). Ethical defaults: Which transparency components can increase the effectiveness of default nudges?. *Social Influenc*e, *14*(3-4), 104-116.

Porat, A., & Strahilevitz, L. J. (2014). Personalizing default rules and disclosure with big data. *Michigan Law Review, 112*(8), 1417-1478. https://repository.law.umich.edu/mlr/vol112/iss8/2

Smith, N. C., Goldstein, D. G., & Johnson, E. J. (2013). Choice without awareness: Ethical and policy implications of defaults. *Journal of Public Policy & Marketing*, *32*(2), 159-172. https://doi.org/10.1509/jppm.10.114

Steffel, M., Williams, E. F., & Pogacar, R. (2016). Ethically deployed defaults: Transparency and consumer protection through disclosure and preference articulation. *Journal of Marketing Research, 53*(5), 865-880. https://doi.org/10.1509%2Fjmr.14.0421

Sunstein, C. R. (2013). Impersonal default rules vs. active choices vs. personalized default rules: A triptych. Active Choices vs. Personalized Default Rules: A Triptych (May 19, 2013). https://dx.doi.org/10.2139/ssrn.2171343

Sunstein, C. R. (2016). People prefer system 2 nudges (kind of). *Duke LJ, 66*, 121.

Thaler, R. H., & Benartzi, S. (2004). Save more tomorrow™: Using behavioral economics to increase employee saving. *Journal of Political Economy, 112*(S1), S164-S187. https://doi.org/10.1086/380085

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.

ESSAY 2

Updated Often Enough: How Product Update Frequency Impacts Consumer Choice

Abstract

Updates, changes to a product to fix issues or add features, are ubiquitous in the consumer environment and occur regularly. Research on updates has focused on the responses of existing users, although updates and how frequently they are released may be meaningful for new consumers as well. In four experiments, we explore how new consumers respond to update frequencies, taking reference points into account. In study 1, consumers were more likely to choose the more updated option when they held higher (compared to lower) expectations about how often a product is updated. In study 2, consumers were more (less) likely to choose the more (less) updated option when they learned about a product updated (less) more than others in the same category. In study 3, consumers were more (less) likely to choose the more updated option when the typical update frequency in the product category was higher (lower). In study 4a, consumers in an incentivized study were more (less) likely to choose a real app that was updated more frequently if the update frequency in the category was higher (lower). In study 4b, consumers in an incentivized study demonstrate sensitivity to update frequency specifically when they were less confident about a product topic, when it was less familiar, when the options were seen as more similar, and when they expected to use it only occasionally.

Keywords: Product updates, Reference points, Cues, New users, Newness, Frequency

INTRODUCTION

Product updates, adjustments to existing features or additional features to meet new needs, have become increasingly ubiquitous, particularly in the domain of digital technologies. Smartphones frequently release and promote updates so consumers can utilize new hardware or software features. For instance, Apple has released over 30 updates to the physical iPhone, with one major update per year in the first six years of the product, increasing to two or, more recently, even three updates per year since then. The update releases are often high-profile events that have become central to Apple's marketing strategy (Tungul, 2019). The use of periodic updates in product marketing is widespread across different product categories (Barry, n.d., Yec, 2022)

Updates typically serve a functional purpose, adjusting existing features or adding new features to meet evolving needs and remain competitive in product quality. For example, Lume cube, a portable lighting company, introduced bi-directional light intensity adjustment as an update in response to customer complaints, and touted their intention to satisfy "every critique, criticism, and wish from our current customer base" (LumeCube, 2019). Thus, updates are also leveraged as a marketing tool, using update releases to bolster marketing efforts. Smule, a music app developer, plans their updates to increase consumer engagement and retention (Apple Inc. n.d.). Marketers are often encouraged to release updates, and to time those releases, in line with consumer engagement goals and to attract consumer attention to increase adoption and usage (Kadir 2023).

In recent years, research has documented that consumers are affected by and respond to update-based marketing strategies. Consumers' behavior regarding the products they currently own changes when new product updates are released (Bellezza, Ackerman, and Gino, 2017; Sela

and LeBouf, 2017). Additionally, research suggests that product updates may be particularly influential for new customers (Jung, Peck, Palmeira, and Kim, 2022; Okada, 2006; Radford and Bloch, 2011; Wood and Lynch, 2002). New consumers may be more receptive to messaging about updates, because consumers with prior knowledge about a product are less likely to learn new information about it (Wood and Lynch, 2002) or to see the product as similar to an owned version (Jung, Peck, Palmeira and Kim, 2022; Okada 2006).

In this paper, we explore how the frequency of free updates affects non-owners' preferences and purchase intentions. Consider someone working on a construction project in their home, choosing between measurement devices, A and B. Since the technology behind many of the measurement device's features is continually developing (and different firms seek to differentiate their product as the best option), the person is aware that their present options will undergo changes. But how does the amount of change it will undergo factor into the decision? For instance, will the frequent updates of device A make it seem more up-to-date than other options, or will they make it seem more unwieldy and undeveloped? If the person notices how updated device A is, will they infer that meaningful changes are being made or will they be attracted to a sign that a product is on the cutting edge of development?

THEORETICAL DEVELOPMENT

The effects of product upgrades

One line of research on product updates has primarily focused on the effect of introducing a new version on consumers who already own a version of the product. This research has shown that consumers value their existing product less when a new version is released. Consumers neglect to report losses of owned products and engage in riskier behavior with owned

products when an update is released (Bellezza, Ackerman, and Gino, 2017). Consumers' decisions to upgrade and adopt the new version is driven by a focus on the advantages of the new version, such that consumers may ignore the benefits of their owned products when making comparisons to new versions unless prompted (Sela and LeBoeuf, 2017).

Another line of research suggests that product updates may be particularly influential for new customers (Jung, Peck, Palmeira, and Kim, 2022; Okada, 2006; Radford and Bloch, 2011; Wood and Lynch, 2002). New consumers may be more receptive to and more likely to learn from messaging about updates than those with prior knowledge (Wood and Lynch, 2002). Those without experience using an existing product may also value an update more than those with product experience, as consumers are less interested in an updated version when it is similar to a product they own (Jung, Peck, Palmeira and Kim, 2022; Okada 2006).

Some of the research on product updates has leveraged the idea of loss-aversion at the level of product attributes. In this view, consumers encountering a version of a product will compare the features and attributes of the updated version to the status quo prior version (Okada 2006), treating the attributes and features of the prior version (or of other well-known products in the category) as the reference point (Amaldoss and He, 2018). A theoretical model incorporating attribute-level reference points has been proposed, which makes predictions regarding the timing of product changes to optimize quality perceptions, based on the product attribute levels relative to those of competitors (Sivakumar and Feng, 2019).

In our research, we focus on a related but distinct (and empirical) question: How does the mere frequency of consumer product upgrades impact consumer perceptions among non-owners, and thereby, affect new product adoption?

Frequency of upgrades

Prior research, which has focused on the objective attribute-level changes in new versions of products, predicts that the effect of more vs. less frequent updates will be driven by the content of those updates. For instance, Okada (2006) implies that a more frequently updated product will be more attractive to the degree that the schedule of updates provides enhancements that make the upgrade less similar to the owned version. Relatedly, more frequent updates will be more effective to the degree that the upgrades deliver greater actual or expected subjective value (Sela and LeBouf, 2017).

Conversely, to the degree that more frequent upgrades raise expectations of the cumulative scope and value of those changes that is then contradicted by information or direct experience with small upgrades, more frequent upgrades may be harmful. For example, as consumers perceived an expense as smaller in magnitude when it was separated into small repeated payments (Gourville, 1998), consumers may view changes made in frequent small updates as insubstantial and ignore them.

However, we propose that upgrade frequency itself can impact consumer decisions, over and above the actual content of those upgrades. Prior research suggests that individuals rely on macro-level cues even when objective information about the relevant details is present. For example, individuals found a presentation to be higher quality when they learned its presenter spent more time preparing (compared to less time preparing), even when the content of the presentations was held constant or when individuals learned about the high cost of the machinery used to produce a product (Chinander and Schweitzer, 2003). Similarly, individuals informed that the same physical items (e.g., paintings, poems, and armor) took more time to create judged the quality to be higher, and reported liking the items more (Kruger, Wirtz, Van Boven, and Altermatt, 2002).

Consumers may respond positively to more frequent updates. Products with more frequent updates may consistently be more attractive because of an "input bias" (Chinander and Schweitzer, 2003), such that irrelevant or redundant information nevertheless boosts perceived quality. In particular, consumers may feel that products with more frequent updates are more continuous and may therefore be less intimidated by change than if the updates were less frequent (Ram and Sheth, 1989). More frequent updates may also provide consumers with a sense of novelty and the feeling that their possession will be "up to date" and currently popular (Jie and Li, 2022). In sum, products with more updates may be seen as reflecting greater effort investment and being of higher quality.

Alternately, however, consumers could also respond negatively to more frequent updates. Consumers may value products that have a longer track record. For example, products from brands labeled as older (rather than labeled more recent) were perceived as higher quality (Pecot, Merchant, de Barnier 2022). Similarly, older medications are preferred to newer medications, even after individuals learned that both drugs were equally effective and affordable, despite newer medications potentially reflecting more advanced research and technology (Jie 2020). Furthermore, consumers may ignore more frequent product updates if they feel too small (Gourville, 1998) or if they are motivated by convenience (Giebelhausen, Robinson, Cronin Jr, 2011). More frequent updates might also signal less investment into producing the new version (Kremer and Debo, 2015).

For consumers to consistently prefer either more frequent or less frequent updates assumes that consumers will be able to evaluate and interpret such frequency information. However, consumers often find it difficult to interpret numeric cues and instead rely on contextual cues for evaluability (Hsee, Hastie, and Chen, 2008; Lambregts, Van Den Bergh,

2019; Zikmund-Fisher, 2019). As a result, consumers may interpret updates relative to a reference point, instead of in absolute terms, using the reference point to assess whether the product is being updated sufficiently frequently. Prior research has demonstrated reference dependence in a variety of decisions, such that consumers are especially sensitive to comparisons to a reference point (Holyoak and Mah, 1982; Hsee, 1998, Hsee, Hastie, and Chen, 2008; Suk, Yoon, Lichtenstein, and Song, 2010).

Product upgrades in practice

Consumers can come across information about update frequency in different contexts. Information about update frequency may be made prominent in product advertisements or descriptions (Barry, n.d., Tungul, 2019, Yec, 2022). Alternately, consumers may learn about update frequency by closely inspecting products or reviews of products they are unsure about (Okada, 2006, Wood and Lynch, 2002, Sela and LeBouf, 2017).

When learning information about update frequency, the nature of the information may influence consumer impressions of products. Consumers may retain more information about necessary fixes if they encode the fix as a negative occurrence (Chen, Lurie, 2013, Frank, Chrysochou, and Mitkidis, 2023). However, when used as a marketing tactic, improvements are more likely to be advertised by firms and may be more top of mind. The tangibility of updates may also influence consumer responses. Consumers may find hardware updates more tangible than software updates and conceptualize them as more substantial (Verhagen, Vonkerman, and van Dolen, 2016). Research on spending finds that the physical appearance of payments influenced the feeling that one was spending money; those using cash felt spending more vividly than those using gift certificates (Raghubir and Srivastava, 2008). Hardware updates may also be more physically noticeable than software updates. In a similar vein, paid updates may be more

noticeable than free updates because consumers will be sensitive to costs (Mazar, Plassmann, Robitaille, and Lindner, 2016).

We propose that consumers coming across information about product update frequency will refer to beliefs about the product's category and use these beliefs as a reference point when making judgments about the product. Those who are unfamiliar with a category may seek externally available information about a category to make product judgments. However, consumers will not always incorporate available information about product updates or have relevant knowledge (internally) or information (externally) available. For instance, we expect that consumers who are committed to a selection or feel knowledgeable about a product category will not rely on update frequency cues. Additionally, we do not expect an effect of update frequency to extend to those who are not sufficiently interested in evaluating or choosing a product (e.g. looking at products for others that they do not use) and ignore update frequency cues.

Across 4 studies, we test a heuristic frequency evaluation framework in which consumers have reference-dependent preferences for more frequent product updates. We find evidence that consumers prefer products with more frequent updates than expected, even holding constant the actual frequency, the content of the updates and manipulating the expected frequency.

STUDY 1: PREFERENCES FOR MORE FREQUENTLY UPDATED PRODUCTS

Method

In this study, we test whether preferences for products with varying frequencies of updates differ based on individuals' personal expectations about update frequencies. In particular, we test between an absolute preference account, in which people simply prefer more

updates, and a relative preference account, in which people especially prefer products that are updated more than their personal expectations. This study was preregistered on aspredicted.org: https://aspredicted.org/P8F_2RD, #56923.

Participants. Participants were recruited via Amazon Mechanical Turk to complete a brief study for $0.30. In all studies, incomplete responses, those with duplicate IP addresses, those with duplicate response IDs, and those who failed any attention checks were excluded from analyses, as pre-registered. We collected 200 complete responses (M_{age} = 38.95, SD_{age} = 11.69; 40% female).

Procedure. Study 1 used a single factor design. All participants read about two construction measurement devices, presented side by side (see Figure 3). One device was updated four times per year (also listed as every three months) and the other was updated eight times per year (also listed as every one and a half months). A short list of changes made to the product in the latest update was included at the end of each description. Product update frequency, product names, and the short description of the changes made in the last update were all individually counterbalanced.

FIGURE 3: Example Product Descriptions in Study 1

Consider the following products:

MarkOne:	**FitX**:
MarkOne is a planning and measurement device used to aid in construction projects. To date, MarkOne has released 4 updates to its original product over the last year, about once every three months.	FitX is a planning and measurement device used to aid in construction projects. To date, FitX has released 8 updates to its original product over the last year, about once every one and a half months.
Over the last year, MarkOne has replaced its 3.2Ghz processor chip with a 3.5Ghz chip that is about 10% faster, and has switched from batteries that last 48 hours after a full charge to batteries that will last 60 hours after a full charge. These are the ONLY differences between the product and its original version.	Over the last year, FitX has replaced its Spectra visual system with the EX system that provides a 10% more accurate reading, and has switched from sensors with a range of 500 feet to sensors with a range of 1250 feet. These are the ONLY differences between the product and its original version.

Then, participants were asked to imagine that they were selecting a device for an upcoming construction project and that the two products were the same price. They then chose between the two products they had read about. To measure expectations about update frequency, participants estimated how many updates they would typically expect each year, in the category of construction measurement devices.

Participants also answered how often they subjectively felt each device was updated each year on a five point scale (from 1=not at all often to 5=extremely often) and listed how many products they had read about in the study as attention checks. Demographics were collected at the end of the study.

Results.

The name of the more updated option and presentation order of the more updated option did not significantly influence choices. The description involving upgrading the Spectra visual system resulted in more choice of that option than the description of the chip upgrade (β=0.246, SE=0.068, p<.001). In the subsequent analyses, we collapse across the counterbalanced factors.

Overall, consumers were more likely to choose the more frequently updated option (59%) than the less frequently updated option (41%; χ^2 = 6.125, df = 1, p = .013). To test between the absolute and relative accounts, we regressed choice of the more updated product on personal expectations about the typical update frequency of the product category. We find that higher expectations about typical update frequency significantly predicted choice, such that people with higher expectations were more likely to choose the more updated product (β =0.040, SE=0.013, p = .002). This result suggests that consumers are sensitive to the relative frequency (i.e., the frequency relative to their expectations), not just the absolute frequency.

To better understand this relationship, we categorized participants into three groups based on their expectations about updates: those typically expecting more updates than either option (i.e., more than 8), those expecting a rate of updates between the more and less frequently updated options (i.e., between 4 and 8), and those expecting fewer updates then either option (i.e., less than 4) As shown in Figure 4, participants were most likely to choose the more updated device when they expected more updates than either option had (76%, N=17, χ^2 = 3.76, df = 1, p = .052), still preferred the more updated device when the less updated option was below their reference point ($61\%, \mathrm{N} = 124, \chi^2$ = 5.88, df = 1, p = .015), but were indifferent when both options were updated at a frequency above their reference point ($49\%, \mathrm{N} = 59, \chi^2$ = 0, df = 1, p = 1).

FIGURE 4: Percent of participants choosing the more updated device by expected update frequency

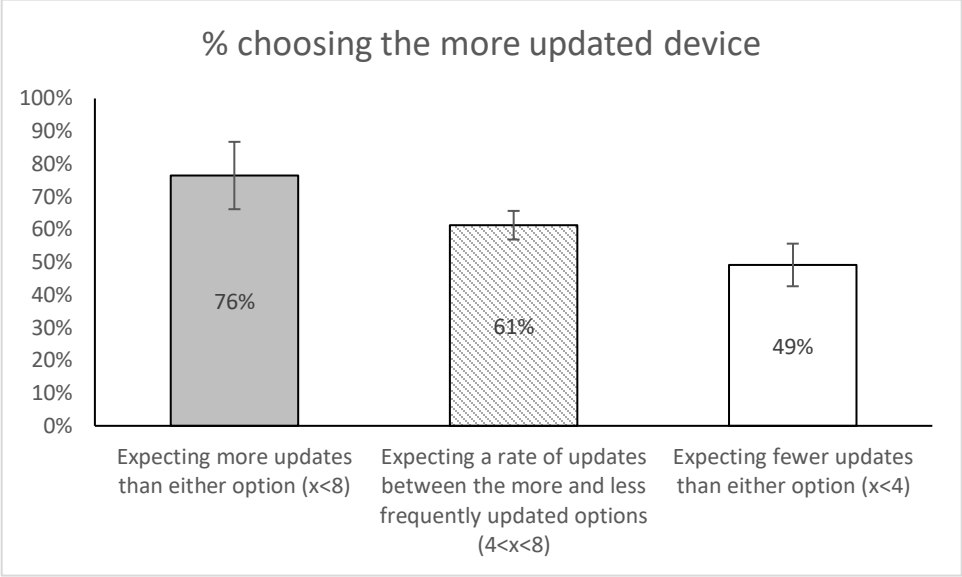


Figure 4. Percent choosing the more updated option by personal reference expectation with standard error bars.

In sum, participants were marginally significantly less likely to choose the less frequently updated product when it was below their reference point (37%), than when it was above their reference point (51%, $\chi^2 = 2.80$, df = 1, p = .094).

The correlational findings from Study 1 suggest that while people prefer products that are updated more frequently, this is preference may be relative to their expectations regarding update frequency. Specifically, holding constant the actual frequency of updates, people were more likely to choose the more frequently updated product when the update frequency for the alternative product was below their expectation level.

STUDY 2: PREFERENCE FOR FREQUENT UPDATES COMPARED TO A FIXED REFERENCE POINT

In this study, we test whether a choice between products with different update frequencies depends on frequency relative to a reference point. Additionally, we explore

consumer inferences that may underlie the preference for more frequently updated products, including individuals' beliefs about firms, product quality, production effort, and interest in social comparison. We also test for the robustness of our findings to physical vs. virtual products. This study was preregistered on aspredicted.org: https://aspredicted.org/blind.php?x=un5ir3, #26315.

Methods.

Participants. Participants were recruited via Amazon Mechanical Turk to complete the experiment for $0.75. Incomplete responses, those with duplicate IP addresses, those with duplicate response IDs, and those who failed any attention checks were excluded from analyses. We collected 202 complete responses (M_{age} = 36.21, SD_{age} = 11.51; 44% female).

Procedure. A 3 (update frequency) x 2 (physical or virtual product) between subjects design was used in this study. Participants read about a target product that had released its 8th update. The target product was randomly assigned to be either physical, a portable LED lighting device, or virtual, a music app. Participants learned how frequently the product was updated, described as the time taken to release its latest version: either six months ago (every six months, twice a year), three months ago (every three months, four times per year), or 1 month ago (every month, 12 times per year). A fixed reference update frequency was also provided: individuals learned that products in the assigned category were typically updated four times a year (every three months). As a result, the target product was updated either more frequently than the reference level, at the same frequency as the reference level, or less frequently than the reference level. As an attention check, participants were also asked to list the update frequency of the product they saw before continuing the study.

For example, participants whose target product was a music app updated every month read:

> "*Version 8 of Sense X has just been released. Sense X is a music app that lets people record and share songs and videos. Typically, a company that produces music apps will regularly do beta-testing or run focus groups to come up with new features to work into the product or ways to fix anything that is not working properly. Once any such changes are decided on and implemented, the final version will be released to the market for consumers as an updated product. A standard music app is updated four times a year, every three months. The firm that produces Sense X typically releases an update about* <u>*every month*</u>*. There was one month between Sense X Version 8 was released and its prior version.*"

After reading about the target product, participants evaluated its quality, effort required to produce it, and the firm producing it. Quality measures included individual responses to each of three questions: the expected overall quality of the product (from 1=very low to 5=very high), how it compared to its competitors (from 1=much worse to 5=much better), and how up-to-date the product seemed to be (1=not at all to 4=very). An overall quality score was made up from the average of the standardized responses to the three quality questions.

Effort measures included individual responses to each of two questions: how much effort it took to produce the product overall (from 1=none at all to 5=lots) and how much the producer had invested in product development (from 1=none at all to 5=very much). An overall effort score was made up from the average of the standardized responses to the two effort questions. The order in which participants evaluated product quality and effort was counterbalanced.

After evaluating quality and effort to produce the target product, participants answered four questions about the firm producing the target product: expected overall quality of management (1=very bad to 5=very good), expected innovativeness (1=not at all to 5=very much), expected long term investment (from 1=none at all to 5=very much), and the firm's capacity to attract people (from 1=very bad, to 5=very good). An overall firm perception score was made up from the average of the standardized responses to the four firm perception questions.

Then, participants made a choice between products. Participants were asked to imagine they had planned to buy a product (a music app or portable light, based on assignment) and had narrowed down their options to three items that were similar in their price, appearance, and functionality. Participants chose between three options: the target product (updated every 1, 3, or 6 months based on randomly assigned condition), a product that was updated every two months, and a product that was updated every four months.

After choosing one of the three products, participants answered three questions about how important social comparison was to them when making a choice. Participants were reminded of the product they had selected when answering these questions. These measures asked to rate the importance that one's choice was acceptable to one's peers, similar to one's peers, or the best among one's peers (from 1=low to 5=high).

As additional attention checks, participants were asked how much time passed between versions of the product (from 1="none at all," to 5="very much") and how many products they had read about. We also asked participants how convenient they thought it was to own the product (from 1="not at all," to 5="extremely").

Results.

Choice of the target product differed significantly based on the frequency of updates (χ^2= 64.43, df = 2, p < .001). When the target product was updated twelve times per year, more than the reference frequency, consumers were most likely (93%) to choose the target product. By contrast, consumers were least likely (27%) to choose the target product when it was updated twice per year, less than the reference frequency. When the product was updated at the target frequency, 57% of participants chose the target product, an intermediate level.

Logistic regression was used to predict the probability of choosing the target product over the two other options, based on whether it was updated more or less frequently than the reference frequency. When the target product was updated more often than the reference frequency, consumers were 91% more likely to choose it than those who did not see a target product updated more than the typical frequency ($\beta = 2.27, SE = 0.53$, p < .001). People who viewed a target product that was updated less than the typical frequency were 22% less likely to choose it than those who did not see a target product updated less than the typical frequency ($\beta = -1.29, SE = 0.37$, p < 0.001).

We also tested whether likelihood of choosing the target product based on reference frequency differed for individuals seeing a virtual instead of a physical product. We did not find an interaction between product type and relative frequency of updates on choice ($\beta_{morethanref} = -0.62, SE = 1.08$, p = 0.568; $\beta_{lesshanref} = 0.96, SE = 1.166$, p = 0.243). Those seeing either a physical or virtual product did not differentially prefer the target product. In addition, we find that reference frequency had a stronger effect on choice than the type of product (whether it was physical or virtual): ($\beta_{morethanref} = 2.57,\ SE = 0.36$, p = .0015; $\beta_{lesshanref} = -2.01, SE =$ .64, p = .0018; $\beta_{virtualtype} = 0.47,\ SE = .54$, p = .38).

Logistic regression was used to test the interaction between the importance of social comparison and relative frequency of updates on choice of the target product. We find a stronger interaction between finding social comparison important and seeing a product updated less than the reference point than between finding social comparison important and seeing a product updated more than the reference point ($\beta = 0.44, SE = 0.66$, p = 0.509). Those who viewed a product updated less than the reference point were 42% ($\beta = 0.96, SE = 0.46$, p = 0.035) more likely to choose the target product (than those who did not see the product updated less than the reference point) with every unit increase in the importance of social comparison.

We also collected participant impressions of the firms producing the stimuli products, product quality, and production effort. We find that firm perceptions were significantly predicted by seeing a target product updated more than the reference frequency (F(1,200)=17.43, p=<0.001; β =0.54, SE =0.13, p<.001) or less than the reference frequency (F(1,200)=43.33, p =<.001; β =-0.79, SE =0.12, p<.001). Firm perceptions significantly predicted choice of the target product (F(1,200)=71.94, p<0.001; β =0.28, SE =0.03, p<0.001). Reference frequency significantly predicted choice of the target product after controlling for firm perceptions ($\beta_{morethanref}$=0.33, SE=0.067, p<0.001; $\beta_{lessthanref}$=-0.162, SE =0.070, p=0.021), suggesting partial mediation. Approximately 22% of the effect of a low reference level on choice and 32% of the effect of a high reference level on choice was accounted for by firm perceptions. The indirect effect was tested using a bootstrap approach with 5000 simulations, using the mediation package in R (Tingley, Yamamoto, Hirose, Keele, Imai, 2014). The indirect effects were significant ($\beta_{morethanref}$=0.118, p<0.001; $\beta_{lessthanref}$=-0.163, p<0.001).

Seeing a target product updated more than the reference frequency also significantly predicted product quality ratings (F(1,200)=25.95, p=<0.001; β =0.66, SE =0.13, p<.001)

Seeing a target product updated less than the reference frequency significantly predicted product quality ratings (F(1,200)=82.79, p=<.001; β =-1.04, SE =0.11, p<.001). Product quality ratings significantly predicted choice of the target product (F(1,200)=72.82, p<.001; β =0.28, SE =0.03, p<.001). Reference frequency significantly predicted choice of the target product after controlling for product quality ratings ($\beta_{morethanref}$=0.39, SE =0.06, p<.001; $\beta_{lessthanref}$=-0.30, SE =0.07, p<.001), suggesting partial mediation. Approximately 26% of the effect of the low reference point and 40% of the effect of the high reference point on choice was accounted for by product quality ratings. The indirect effect was tested using a bootstrap approach with 5000 simulations, using the mediation package in R (Tingley, Yamamoto, Hirose, Keele, Imai, 2014). The indirect effects were significant ($\beta_{morethanref}$= 0.14, p<.001; $\beta_{lessthanref}$=-0.20, p<.001).

Finally, seeing a target product updated more than the reference frequency significantly predicted perceived effort to produce the product (F(1,200)=10.08, p=.002; β =0.42, SE =0.13, p=.002). Seeing a target product updated less than the reference frequency significantly predicted perceived effort to produce the product (F(1,200)=29.32, p=<.001; β =-0.67, SE=0.12, p<.001). Perceived effort to produce the product significantly predicted choice of the target product (F(1,200)=21.38, p<.001; β =0.17, SE=0.04, p<.001). Reference frequency significantly predicted choice of the target product after controlling for perceived effort to produce the product ($\beta_{morethanref}$=0.48, SE=0.06, p<.001; $\beta_{lessthanref}$=-0.44, SE=0.07, p<.001) , suggesting partial mediation. Approximately 9% of the effect of the low reference point and 11% of the effect of the high reference point on choice was accounted for by perceived effort to produce the product. The indirect effect was tested using a bootstrap approach with 5000 simulations, using the

mediation package in R (Tingley, Yamamoto, Hirose, Keele, Imai, 2014). The indirect effect was significant ($\beta_{morethanref}$ = 0.047, p=0.002; $\beta_{lessthanref}$ =-0.057, p=0.011).

STUDY 3: MANIPULATING REFERENCE FREQUENCY

In the prior studies, we found a preference for more frequent updates, relative to consumers' own reference frequencies (Study 1) and to a fixed reference frequency (Study 2). These results suggest but do not directly test for sensitivity specifically to relative frequency. In Study 3, we test the effect of manipulating the reference frequency on choices and evaluations of products, holding constant the update frequency of the target product. Participants learned how frequently construction measurement devices typically underwent change when considering a choice between devices and evaluating them. This study was preregistered on aspredicted.org: https://aspredicted.org/blind.php?x=qt65iu, #14208.

Method

Participants. Participants were recruited via Amazon Mechanical Turk to complete the experiment for $0.40. Incomplete responses, those with duplicate IP addresses, those with duplicate response IDs, and those who failed any attention checks were excluded from analyses. We collected 480 complete responses (M_{age} = 37.34, SD_{age} = 11.72; 48% female).

Procedure. This study used a 3 condition between-subjects design. To test whether reference points influenced choices between products updated at different rates, participants were randomly assigned to either read that products in the category of construction measurement devices were typically updated two times per year, six times per year, or ten times per year. As an attention check, participants had to list how many times construction measurement devices were typically updated before moving on to the rest of the study.

Participants then read about two construction measurement devices, presented side by side. One device was updated four times per year (also listed as every three months) and the other was updated eight times per year (also listed as every one and a half months). The order, name and description of the options were all counterbalanced. A short list of changes that had been made to the product in the latest update was included at the end of each description. Product update frequency, product names, and the short description of the changes made in the last update were presented were all counterbalanced separately. For example, participants assigned to the 2 updates per year reference frequency viewed:

"Thank you for participating in this survey. On the following pages, you will read about two different construction measurement devices and then make judgments about them.

Please note that products in this category are typically updated 2 times a year."

Then, participants saw the same construction measurement stimuli from Study 1 on the next page and read:

"Imagine that you have to choose a measurement device for an upcoming construction project. If your budget allows you to select either product you read about, which would you choose to purchase?"

Participants then made a choice between "The device that releases about 4 updates a year, [counterbalanced name]" and "The device that releases about 8 updates a year, [counterbalanced name]."

For those learning that the typical frequency of updates was two times per year, both options they were choosing between would be updated more than the reference frequency. For those learning that the typical frequency of updates was six times per year, one item they were choosing between would be updated more frequently than the reference frequency and the other

item would be updated less frequently than the reference frequency. For those learning that the typical frequency of updates was ten times per year, both items they were choosing between would be updated less than the reference frequency. Next, participants were asked to choose between the two items they read about, imagining that they were selecting a device for an upcoming construction project and that either product was within budget.

To explore whether reference frequencies influenced judgments of product quality and effort evaluations of products changed at different rates, participants answered the same questions about the expected quality and the expected effort required to produce each device after making a choice as in Study 2. Prior to rating quality and effort, participants were reminded of the description of the product they were evaluating.

As additional attention checks, participants answered how often they thought each of the products was updated each year on a 5 point scale (from 1, "Not at all often" to 5, "Extremely often") and listed how many products they had read about in the study. To explore possible reasoning for update frequency preference participants were also asked how many updates they typically expected from a product like a measurement device (on a slider scale from "0," indicating none to "24," indicating 24 and above), and how convenient it would be to own each of the products on a 5 point scale (from 1, "Not at all" to 5, "Extremely").

Results.

The name (χ^2= 0.13, df = 1, p = .716), description (χ^2= 1.07, df = 1, p = .300) and order of presenting (χ^2= 0.0002, df = 1, p = .987) the more updated option did not have a significant effect on choice. We collapse across these differences in our subsequent analyses.

We observed that choice of the more frequently updated product was significantly more likely with higher reference frequencies ($\beta_{inbetweenoptions,6}$=0.13, SE=0.06, p=.017;

$\beta_{morethanoptions,10}$==0.19, SE=0.05, p<.001). Participants choices differed based on the reference level (χ^2 = 12.87, df = 2, p = .002). They favored the more frequently updated product (63%) when both options were updated less frequently than the reference frequency, somewhat less (57%) when only one option was updated more frequently than the reference frequency, and the least (41%) when both options were updated more frequently than the reference frequency. In particular, choices of the more frequently updated option were significantly lower when the reference point was less frequent than either option (twice per year), compared to when the reference point was between the two options (6 times per year; χ^2 = 5.04, df = 1, p = .025) or was more frequent than either option (10 times per year; χ^2 = 11.50, df = 1, p < .001).

A factor analysis confirmed that the three perceived quality questions and the two perceived effort questions loaded on separate factors (see Online Appendix). Accordingly, we combined the three quality questions into an index of perceived quality (alpha = 0.920) and the two effort questions into an index of perceived effort (alpha = 0.900). Quality scores and effort scores were distinguishable but highly positively correlated, r=.822, p<.001.

The difference in general perceptions of quality between the two items did not differ significantly based on reference frequency (F(2,477)=0.21, p=.813; $\beta_{refbetweenoptions}$=-0.03, $\beta_{refaboveoptions}$=-0.09). This was the case for individual measures of overall quality (F(2,477)=0.07, p=.938; $\beta_{refbetweenoptions}$=0.049, $\beta_{refaboveoptions}$=0.045), how well the product would compare to peers (F(2,477)=0.61, p=.544; $\beta_{refbetweenoptions}$=-0.159, $\beta_{refaboveoptions}$=-0.153), and how up-to-date the devices seemed (F(2,477)=0.877, p=.417; $\beta_{refbetweenoptions}$=0.018, $\beta_{refaboveoptions}$=-0.168).

The difference in general perceptions of effort to produce between the two items did not differ significantly based on reference frequency (F(2,477)=0.324, p=.723; $\beta_{refbetweenoptions}$=-

0.116, $\beta_{refaboveoptions}$=-0.038). This was the case for individual measures of overall effort needed to produce the products (F(2,477)=0.641, p=.528; $\beta_{refbetweenoptions}$=-0.168, $\beta_{refaboveoptions}$=-0.043) and difference in perceived investment from the producer (F(2,477)=0.088, p=.916; $\beta_{refbetweenoptions}$=-0.063, $\beta_{refaboveoptions}$=-0.034).

The difference in perceived quality between items was marginally significantly correlated with choice of the more frequently updated product (r=-0.086, p=.059); as the difference in perceived quality between products increased, choice of the more updated product decreased slightly. Differences in perceived effort to produce the two products and choice of the more frequently updated product were not significantly correlated (r=-0.062, p=.173). Additionally, the difference in perceived convenience of the products did not vary significantly based on the reference frequency of updates (F(2,477)=0.259, p=.772; $\beta_{refbetweenoptions}$=0.072, $\beta_{refaboveoptions}$=0.102).

In Experiment 3, we find that, holding actual update frequency constant, differences in the reference level of updates affects choices of a more frequently updated product over a less frequently updated product, consistent with a relative assessment of update frequency. In Experiment 3, we tested individuals' preferences between hypothetical products in an unfamiliar product category. We did not find evidence that product evaluations were affected by the reference frequencies, suggesting that consumers made their choices based on their direct preferences for relative frequencies, rather than because of quality or effort inferences from those updates.

STUDY 4A: THE EFFECT OF MANIPULATED REFERENCE FREQUENCY ON CONSEQUENTIAL CHOICES

We explore people's preferences for real products undergoing different frequencies of change when a reference frequency is present. Participants learned how frequently iPhone apps typically underwent change when choosing between them. Categories of apps used in the studies were determined in a pretest. In the pretest, individuals viewed 20 apps in 7 categories in random order and rated how interested they would be in downloading each one on a 3 point scale if they did not already own it. Stargazing and fitness were selected as participants had the most interest in downloading apps in these categories. No process measures were collected in study 4a. This study was preregistered on aspredicted.org: https://aspredicted.org/X83_85P, #79443.

Method.

Participants. Participants were recruited via Amazon Mechanical Turk. So that participants taking part in the survey would have access to the specific apps, participants first completed a screening survey for $0.15. Those who indicated that they used an iPhone in the screener survey were invited to take the main experiment for $0.85. We collected 602 complete responses (M_{age} = 36.80, SD_{age} = 11.77; 60% female).

Procedure. This study used a 2 condition between-subjects design, varying the reference frequency. Participants were randomly assigned to see an estimated frequency at which stargazing apps were updated that was less frequent (0.2/yr) or more frequent (4/yr) than the individual stargazing apps they would read about (see stimuli in Online Appendix).
The estimated frequencies for updates in this category of app was determined by taking the average of update frequencies of at least five similar apps in the Apple app store. The high estimated reference frequency (4/year) was the overall average frequency of stargazing apps at the time of the study, while the low estimated reference frequency (0.2/yr) was not.

Participants then made a consequential choice between two paid iPhone applications. Participants were shown an image and brief description of each app, side by side in counterbalanced order, as shown in Figure 5. We used real apps from the Apple app store. One stargazing app was updated roughly every month (0.9/yr), and the other was updated every four months (three times a year). As a result, both apps were updated more frequently than the reference level in one condition, and both were updated less frequently than the reference level in the other condition.

The study instructions included information that highlighted that the choice could be realized. Prior to beginning the study, participants learned that there were two parts of the survey and the compensation for taking each part. Participants would receive $0.15 for participating in part 1, a screener that would ask what type of smart phone one used among decoy questions (e.g. what web browser do you use, which streaming service do you prefer, among others). Participants would receive $0.85 for participating in part 2, a choice task. Lastly, to emphasize that one could receive an item they chose, participants learned that 5 participants would be randomly selected to win the item they selected in part 2.

Participants were reminded of this incentive before starting the choice task. All participants taking part 2 read:

"Thank you for participating in Part 2 of this survey. In this part, you will read about two apps, evaluate each app, and choose one you would prefer to download. Please do not look up the information in the study, as it will invalidate the results.
Please choose carefully. <u>You will have a chance to win the app you select in the survey</u>."

As attention checks, participants answered how often they thought each of the products was updated each year on a 5 point scale (from 1, “Not at all often” to 5, “Extremely often”) and listed how often apps in the category were updated.

FIGURE 5: Example consequential choice screen

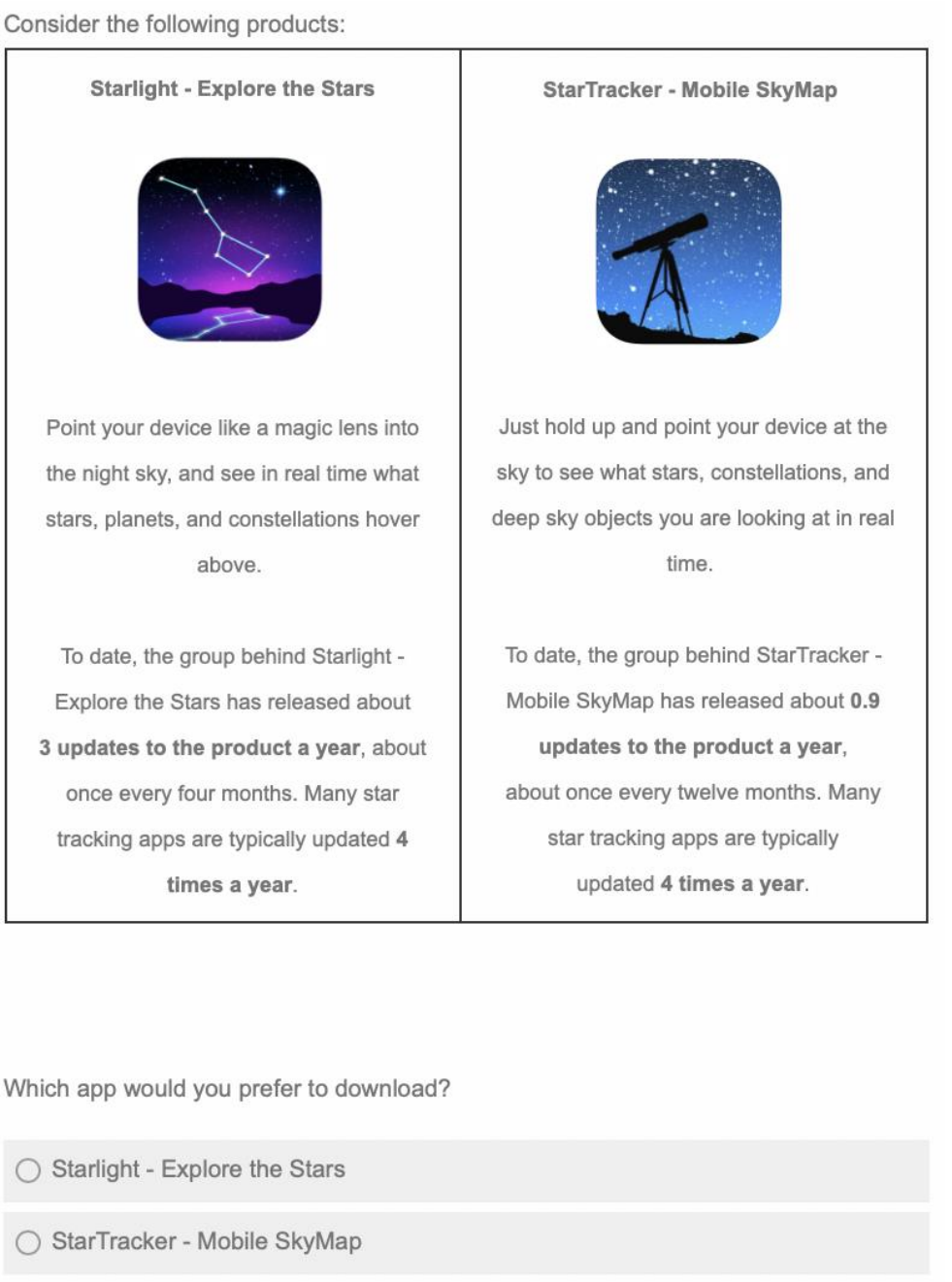


Figure 5. An example of the choice task in Study 4a. Descriptions and update frequencies were taken from the Apple app store. Apps were shown in counterbalanced order.

Results.

We ran a logistic regression predicting choice of the more frequently updated option based on the reference point condition, controlling for order of presentation. As predicted, we find a positive coefficient of reference point frequency (β=0.147, SE=0.053, p=.006). Participants were more likely to choose the more frequently updated app when both options were below the high reference update frequency (83%) than when the same two options were both above the low reference updated frequency (74%; χ^2 = 7.147, p = .004).

In Study 4a, we replicate the effect of manipulating the reference frequency on consequential choices between real products, holding the actual update frequencies of the real products constant. Consumers were more likely to choose a more frequently updated product when both options were below the reference level, compared to when both options were above the reference level. Next, we test the generality of this finding to different types of products.

STUDY 4B: PRODUCT TYPE MODERATES THE EFFECT OF MANIPULATED REFERENCE FREQUENCY ON CONSEQUENTIAL CHOICES

We explore product type as a potential boundary condition for the effect of the reference frequency on consumer choices. In this study, we used the same design as Study 4a, with two different categories of iPhone apps. We find that manipulating relative update frequency impacts choices between products in a low-expertise product category (in which individuals are less confident about their choice, find options more similar, are less familiar with product content, and intend to use the app less frequently) but not in a higher expertise product category. This study was preregistered on aspredicted.org: https://aspredicted.org/blind.php?x=jq7b9s, #33983. Two replications of this study were pre-registered and conducted: (1) using different apps from

study 4b, without bolded frequencies, and a $0.10 screener payment, #29101 https://aspredicted.org/blind.php?x=9e2sg4, and (2) using the same apps as in study 4b, without bolded updated frequencies, #30492 https://aspredicted.org/blind.php?x=9qk9pg. The results are reported in the appendix.

Method.

Participants. Participants were recruited via Amazon Mechanical Turk. So that participants taking part in the survey would have access to the specific apps, participants first completed a screening survey for $0.15. Those who indicated that they used an iPhone in the screener survey were invited to take the main experiment for $0.85. We collected 203 complete responses (M_{age} = 36.39, SD_{age} = 12.57; 57% female).

Calibration Test. In a separate calibration test (Mturk, N=94), we tested perceptions of two different types of apps that we used in the study: stargazing apps and fitness apps. The stargazing apps identified stars and other objects in the sky and the fitness apps allowed users to put exercises together to create workouts (see the stimuli used in Figure 6, and Supplement 3 of the Appendix). We find that individuals found that the stargazing apps were seen as significantly more similar (ps<.003), less familiar in terms of product content (ps<.001) and had less frequent usage intentions (ps<.02) than the fitness apps (see Online Appendix for details).

Figure 6. Calibration test app stimuli

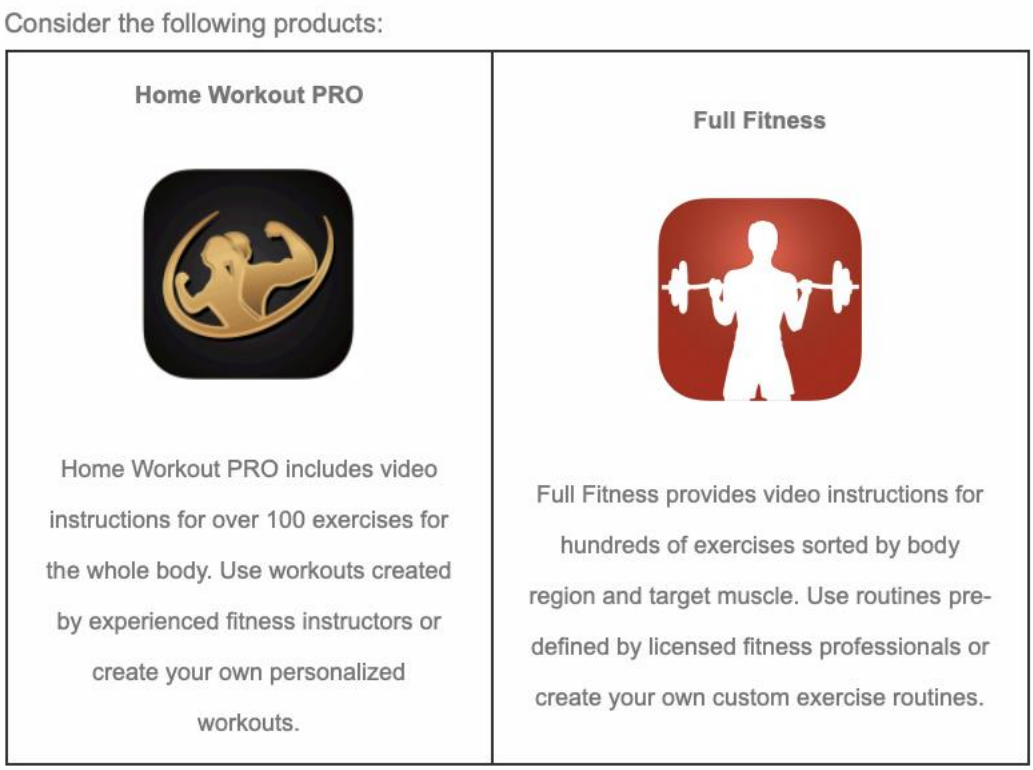
Consider the following products:

Starlight: Explore the Stars	StarTracker - Mobile SkyMap
	
Point your device like a magic lens into the night sky and see in real time what stars, planets, and constellations hover above. Stargazing has never been so easy!	Just hold up and point your device at the sky to see what stars, constellations, and deep sky objects you are looking at in real time.

Consider the following products:

Fitness Buddy: Gym Workout Log	Streaks Workout
	
Be stronger. Be leaner. Be the best you. Build your own workout routines. Find out why gym goers are switching over to Fitness Buddy!	The personal trainer that you actually want to use. Exercise anywhere! Customize which exercises you want to do, making it great for all ages and abilities.

Figure 6. Continued.



Main Study Procedure. This study used a 2 (between: low vs. high reference frequency of updates) x 2 (within: stargazing vs. fitness app types) mixed design. Participants were randomly assigned to read one of the two estimated reference frequencies for the stargazing apps and one of the two estimated reference frequencies for the fitness app, determined by taking the average of update frequencies of at least five similar apps in the Apple app store. In one condition, participants were shown the high reference frequency for both apps: 4.5/year for the star-gazing apps, which was the overall average reference frequency for stargazing apps at the time of the study, and 14/year for the fitness apps, which was not the overall average. In the other condition, participants were shown the low reference frequency for both apps: 0.2/year for the star-gazing apps, which was not the overall average, and 6.5/year for the fitness apps, which was the overall average for fitness apps at the time of the study.

Participants then made two choices between pairs of paid iPhone applications, in counterbalanced choice and presentation order. One choice set included the same two stargazing apps as in Study 4a and the other included two fitness apps. As in Study 4a, real apps from the Apple app store were used along with their descriptions and update frequencies. One stargazing app was updated 3 times a year, and the other was updated 0.9 times a year. One fitness app was updated 13 times a year, and the other was updated 8 times a year. As a result, the apps' update frequencies were all below the relevant reference level in the high reference condition, and the apps' update frequencies were all above the relevant reference level in the low reference condition.

To explore whether reference frequencies influenced judgments of product quality and effort investment in the products, participants answered the same questions about the expected quality and the expected effort required to produce each device after making a choice, as in study 2. Prior to rating quality and effort, participants were reminded of the description of the product they were evaluating.

As an attention check, participants answered how often they thought each of the products was updated each year on a 5 point scale (from 1="Not at all often" to 5="Extremely often").

Results.

Replicating Study 4a, participants were more likely to choose the more frequently updated stargazing app (89%) in the high-reference-level condition, when both apps' update frequencies were below the reference level, than in the low-reference-level condition (78%; χ^2 = 4.243, p = .039), when both apps' update frequencies were below the reference level. By contrast, choice of the more frequently updated fitness app did not differ significantly between the high and low reference frequency conditions (75% vs. 73%; χ^2 = 0.021, p = .885). While the

design of this study did not allow us to test an interaction between the reference-level effect and the product type, our results demonstrate the robustness of the effect for star-gazing apps, a low expertise category, and a failure to generalize to fitness apps, a higher expertise category.

Consistent with the results of a factor analysis (see Online Appendix), we combined the three quality questions into an index of perceived quality separately for each app type (Stargazing: alpha = 0.87, Fitness: alpha = 0.84) and the two effort questions into an index of perceived effort separately for each app type (Stargazing: alpha = 0.75, Fitness: alpha = 0.69). Quality scores and effort scores were distinguishable but highly positively correlated (Stargazing: r=.677, p<.001, Fitness: r=.689, p<.001).

Participants perceived a greater difference in quality between more and less frequently updated stargazing apps in the high-reference-frequency condition, where both options were below the reference level, than in the low-reference-frequency condition, where both options were above the reference level (F(1,201)=10.73, p=0.001, $\beta = 0.403, p = .001$). Likewise, participants perceived a greater difference in quality between more and less frequently updated fitness apps in the high-reference-frequency condition than in the low-reference-frequency condition (F(1, 201)=4.78, p=0.029, $\beta = 0.266, p = .029$).

Participants also perceived a greater difference in effort invested between more and less frequently updated stargazing apps in the high-reference-frequency condition, where both options were below the reference level, than in the low-reference-frequency condition, where both options were above the reference level (F(1, 201)=6.033, p=0.015, $\beta = 0.306, p = .015$). This was also the case for fitness apps based on reference frequency (F(1, 201)=6.014, p=0.015, $\beta = 0.298, p = .015$).

In a logistic regression predicting the probability of choosing the more frequently updated stargazing app, perceived quality difference predicted choice ($\beta = 1.946, SE = 0.323$, p<.001) while reference frequency was not a significant predictor, controlling for perceived quality (β=0.454, SE=0.452, p=.315). Likewise, perceived effort difference significantly predicted choice of star-gazing app ($\beta = .657, SE = .242$, p=.007) but reference frequency was not a significant predictor, controlling for perceived effort (β=0.589, SE=0.423, p=.163).

While reference frequency did not affect choices of the more frequently updated fitness app, both perceived quality ($\beta = 1.946, SE = 0.323$, 95% CI=[1.361, 2.633], p<.001) and perceived effort separately predicted probability of choosing the more frequently updated fitness app, controlling for reference frequency. ($\beta = 1.457, SE = 0.281$, 95% CI=[0.943, 2.049], p<.001).

We tested whether differences in perceived quality or production effort between the more and less updated stargazing apps mediated the effect of choosing the more updated stargazing app given reference frequency. Reference frequency significantly predicted differences in perceived product quality (β =0.40, SE =0.12, p=.001). Difference in perceived product quality significantly predicted choice of the more updated stargazing app (β =0.19, SE=0.03, p<.001). Reference frequency did not predict choice of the more updated stargazing app after controlling for differences in perceived quality ($\beta_{quality}$=0.18, SE=0.03, p<.001; $\beta_{reference}$=0.04, SE=0.05, p=.365). Approximately 62% of the effect of reference frequency on choice was accounted for by perceived effort to produce the product. The indirect effect was tested using a bootstrap approach with 5000 simulations, using the mediation package in R (Tingley, Yamamoto, Hirose, Keele, Imai, 2014). The indirect effect was significant ($\beta_{reference}$= 0.07, p=.001).

Reference frequency significantly predicted differences in perceived production effort (β =0.31, SE =0.12, p=.015). Difference in perceived production effort significantly predicted choice of the more updated stargazing app (β =0.13, SE =0.03, p<.001). Reference frequency did not predict choice of the more updated stargazing app after controlling for differences in perceived production effort (β_{effort}=0.12, SE=0.03, p<.001; $\beta_{reference}$=0.08, SE=0.05, p=.117). Approximately 32% of the effect of reference frequency on choice was accounted for by perceived effort to produce the product. The indirect effect was tested using a bootstrap approach with 5000 simulations, using the mediation package in R (Tingley, Yamamoto, Hirose, Keele, Imai, 2014). The indirect effect was significant ($\beta_{reference}$= 0.04, p=.014).

In Study 4b, we replicate the impact of higher manipulated reference update frequency on choices of a more frequently updated stargazer app, but the effect did not generalize to a different app category, in which participants had more experience. In particular, in a separate calibration test, participants were more confident, perceived larger differences, were more familiar with and had higher intentions to use the fitness apps, for which the effect was not observed. Nevertheless, participants' choices were similarly related to perceived quality and effort investment of the apps. In sum, these results suggest that consumers may base their quality and effort investment perceptions on update frequency relative to a reference level, impacting their choices, when making decisions in less well-known categories, but may incorporate other, potentially more directly relevant, factors when making decisions in more well-known categories.

GENERAL DISCUSSION

Findings from these studies suggest that new consumers may utilize update frequency to evaluate and decide between products. In study 1, we observe consumer preferences for updates when no external reference point for update frequency is provided. We asked participants to choose between two products in an unfamiliar category updated at different frequencies and found that a sizeable portion of consumers were not attracted to the more updated option when there was no external reference point given, suggesting that consumers may not simply be interested in the most updated option. However, consumers also reported how often they thought a product in the unfamiliar category was updated. We found that as personal expectations about how often a typical product was updated increased, people were more likely to select the more updated option, suggesting that consumers may compare update frequency to an internal reference point.

In study 2, information about how often a product was typically updated was provided to participants. In the study, participants chose between more familiar products like portable lights and music apps. We found that individuals were more likely to choose a product if the rate it was updated exceeded (rather than falling short of) the rate they learned was typical of the category, suggesting that information about update frequency is utilized as a threshold. This effect is robust to whether one is considering a physical product (like the portable light) or a virtual product (like a music app). However, the stimuli used in the study assume that updates to a physical item will involve physical changes and updates to a virtual item will involve virtual changes. Additionally, we find that consumer interest in social comparison only encouraged choice when one saw a product updated less frequently than the typical rate of updates, suggesting social comparison is not a consistent driver of interest in more frequent updates.

In study 3, participants chose between two unfamiliar products and we manipulated whether the typical rate of updates for a category fell below, between or above the options. Consumers were most interested in having the most updated option when they learned that products in the category were typically updated more than their options. They were less interested in the most updated option if their options exceeded or fell on either side of the typical frequency, supporting that simply increasing product updates will not be attractive to consumers in all contexts.

Finally, in studies 4a and 4b, we test the influence of product update frequency with real products in incentive compatible experiments. Participants chose between real smart phone apps and would receive one of the products they selected in the experiment. By comparing products with different attributes, we demonstrate boundary conditions for consumer use of update frequency information. Findings from these studies suggest that consumers may rely on relative update frequency when they are less confident about their choice, find options similar, are less familiar with product content, or when they intend to use the product less frequently.

From studies 2, 3, and 4b, we find mixed evidence for consumer interpretations of and inferences from different perceived update frequency. Results from study 2 and 4b suggest that preferences for different rates of update frequency may stem from consumer perceptions of product quality and the effort required to produce the product. Findings from study 2 also suggest that perceptions about the firm producing the product contribute to preferences for different rates of product updates. In study 3, we find that perceptions of quality, production effort, and firm perceptions between products did not significantly differ as relative category updates moved. This could suggest that information about product category updates may not substantially differentiate attributes of products updated at different rates from one another.

Furthermore, in Study 4b, we found that reference level differences impacting quality and effort perceptions did not necessarily translate into an effect on choice, in the case of the fitness apps.

Results from these studies suggest that consumers can use update frequency information, generated internally or provided externally, as a reference point to evaluate products in multiple settings. Taking this into account, it is possible that update frequency reference points can be employed to promote product choice for new customers or those less familiar with a product. For example, firms may present explicit information about category update frequency to hamper interest in competitors releasing more updates. Further research can explore in depth the influence that different forms of updates have on how consumers use update frequency. Research can also shed light on whether information about update frequency remains influential for new consumers when other readily understandable information is also present.

References

Amaldoss, W., & He, C. (2018). Reference-dependent utility, product variety, and price competition. *Management Science, 64*(9), 4302-4316.

Apple Inc. (n.d.), *Smule - Developer Insights - App Store.* Apple Developer. https://developer.apple.com/app-store/smule/.

Barry, K. (n.d.). *Automakers embrace over-the-air updates, but can we trust Digital Car Repair?* Consumer Reports. Retrieved from https://www.consumerreports.org/automotive-technology/automakers-embrace-over-the-air-updates-can-we-trust-digital-car-repair/

Bellezza, S., Ackerman, J. M., & Gino, F. (2017). "Be careless with that!" Availability of product upgrades increases cavalier behavior toward possessions. *Journal of Marketing Research, 54*(5), 768-784.

Chen, Z., & Lurie, N. H. (2013). Temporal contiguity and negativity bias in the impact of online word of mouth. *Journal of Marketing Research, 50*(4), 463-476.

Chinander, K. R., & Schweitzer, M. E. (2003). The input bias: The misuse of input information in judgments of outcomes. *Organizational Behavior and Human Decision Processes, 91*(2), 243-253.

Cunha, Jr, M., & Caldieraro, F. (2009). Sunk-cost effects on purely behavioral investments. *Cognitive Science, 33*(1), 105-113.

Frank, D. A., Chrysochou, P., & Mitkidis, P. (2023). The paradox of technology: Negativity bias in consumer adoption of innovative technologies. *Psychology & Marketing, 40*(3), 554-566.

Giebelhausen, M. D., Robinson, S. G., & Cronin, J. J. (2011). Worth waiting for: increasing satisfaction by making consumers wait. *Journal of the Academy of Marketing Science*, *39*(6), 889-905.

Gourville, J. T. (1998). Pennies-a-day: The effect of temporal reframing on transaction evaluation. *Journal of Consumer Research*, *24*(4), 395-408.

Hamilton, R. W., Ratner, R. K., & Thompson, D. V. (2011). Outpacing others: When consumers value products based on relative usage frequency. *Journal of Consumer Research, 37*(6), 1079-1094.

Hsee, C. K., Hastie, R., & Chen, J. (2008). Hedonomics: Bridging decision research with happiness research. *Perspectives on Psychological Science*, *3*(3), 224-243.

Jie, Y. (2020). Older is better: Consumers prefer older drugs. *Psychology & Marketing*, *37*(11), 1498-1510.

Jung, W., Peck, J., Palmeira, M., & Kim, K. (2022). An Unintended Consequence of Product Upgrades: How Upgrades Can Make Current Consumers Feel Left Behind. *Journal of Marketing Research*, 00222437221078551.

Kadir. (2023, February 16). *When to Release an Update for Your Mobile App*. AppSamurai. https://appsamurai.com/blog/when-to-release-an-update-for-your-mobile-app/.

Kahneman, D. (1979). Prospect theory: An analysis of decisions under risk. *Econometrica*, *47*, 278.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1991). Anomalies: The endowment effect, loss aversion, and status quo bias. *Journal of Economic perspectives*, *5*(1), 193-206.

Kremer, M., & Debo, L. (2016). Inferring quality from wait time. *Management Science*, *62*(10), 3023-3038.

Lembregts, C., & Van Den Bergh, B. (2019). Making each unit count: The role of discretizing units in quantity expressions. *Journal of Consumer Research, 45*(5), 1051-1067.

LumeCube. (2019, October 7). *Lume Cube 2.0: The Long Awaited 2nd Generation*. Lume Cube Inc. https://lumecube.com/blogs/news/lume-cube-2-0.

Mazar, N., Plassmann, H., Robitaille, N., & Lindner, A. (2016). Pain of paying?—A metaphor gone literal: Evidence from neural and behavioral science. *Rotman School of Management Working Paper*, (2901808).

McConnell, J. D. (1968). The price-quality relationship in an experimental setting. *Journal of Marketing Research*, *5*(3), 300-303.

Neumann, N., & Böckenholt, U. (2014). A meta-analysis of loss aversion in product choice. *Journal of Retailing*, *90*(2), 182-197.

Okada, E. M. (2006). Upgrades and new purchases. *Journal of Marketing*, *70*(4), 92-102.

Radford, S. K., & Bloch, P. H. (2011). Linking innovation to design: Consumer responses to visual product newness. *Journal of Product Innovation Management*, *28*(s1), 208-220

Ram, S., & Sheth, J. N. (1989). Consumer resistance to innovations: the marketing problem and its solutions. *Journal of consumer marketing*.

Raghubir, P., & Srivastava, J. (2008). Monopoly money: The effect of payment coupling and form on spending behavior. *Journal of Experimental Psychology: Applied, 14*(3), 213–225.

Schwartz, B., Ward, A., Monterosso, J., Lyubomirsky, S., White, K., & Lehman, D. R. (2002). Maximizing versus satisficing: happiness is a matter of choice. *Journal of personality and social psychology*, *83*(5), 1178.

Sela, A., & LeBoeuf, R. A. (2017). Comparison neglect in upgrade decisions. *Journal of Marketing Research*, *54*(4), 556-571.

Simonson, I., & Tversky, A. (1992). Choice in context: Tradeoff contrast and extremeness aversion. *Journal of marketing research*, *29*(3), 281-295.

Sivakumar, K., & Feng, C. (2019). Patterns of product improvements and customer response. *Journal of Business Research*, *104*, 27-43.

Tungul, Jade. (2019, September). *How Apple's Keynote Formula Keeps Audiences Engaged*. Business Insider. https://www.businessinsider.com/apple-keynote-events-engaging-product-launch-anticipation-2019-9.

Tversky, A., & Simonson, I. (1993). Context-dependent preferences. *Management science*, *39*(10), 1179-1189.

Verhagen, T., Vonkeman, C., & van Dolen, W. (2016). Making online products more tangible: the effect of product presentation formats on product evaluations. *Cyberpsychology, Behavior, and Social Networking*, *19*(7), 460-464.

Wirtz, D., Kruger, J., Altermatt, W., & Van Boven, L. (2004). The effort heuristic. *Journal of Experimental Social Psychology*, *40*, 91-98.

Wood, S. L., & Lynch Jr, J. G. (2002). Prior knowledge and complacency in new product learning. *Journal of Consumer Research*, *29*(3), 416-426.

Yec. (2022, November 8). *Council post: 8 tips for releasing major product updates*. Forbes. Retrieved from https://www.forbes.com/sites/theyec/2022/11/07/8-tips-for-releasing-major-product-updates/?sh=71686b2f2449

Zikmund-Fisher, B. J. (2019). Helping people know whether measurements have good or bad implications: increasing the evaluability of health and science data communications. *Policy Insights from the Behavioral and Brain Sciences*, *6*(1), 29-37.

Appendix A: Essay 2

SUPPLEMENTAL MATERIALS FOR:

Updated Often Enough: How Product Update Frequency Impacts Consumer Choice

Authors: Shweta R. Desiraju, Oleg Urminsky

Please find the most up-to-date version of this document and other supplemental materials at: https://osf.io/t4huj/?view_only=7f59c2b673c74c6989ff0ac65443867c

Supplement 1: Study 3 additional results

Minimum residual factor analysis was conducted for the five questions relating to product quality and effort to produce products. This yielded two factors explaining 83% of the variance for the set of five variables. Factor 1, with high loadings on difference scores of overall perceived quality (0.82), how the product compared to its peers (1.00), and how up-to-date the product seemed (0.60), was labeled perceived quality. Factor 1 explained 48% of the variance. Factor 2, with high loadings on difference scores of overall perceived effort to produce the product (0.63) and perceived investment from the producer (1.00), was labeled perceived effort. Factor 2 explained 35% of the variance. Factor 1, perceived quality, and Factor 2, perceived effort, were highly positively correlated (r=0.82).

Supplement 2: Study 4a additional results

Example app stimuli:

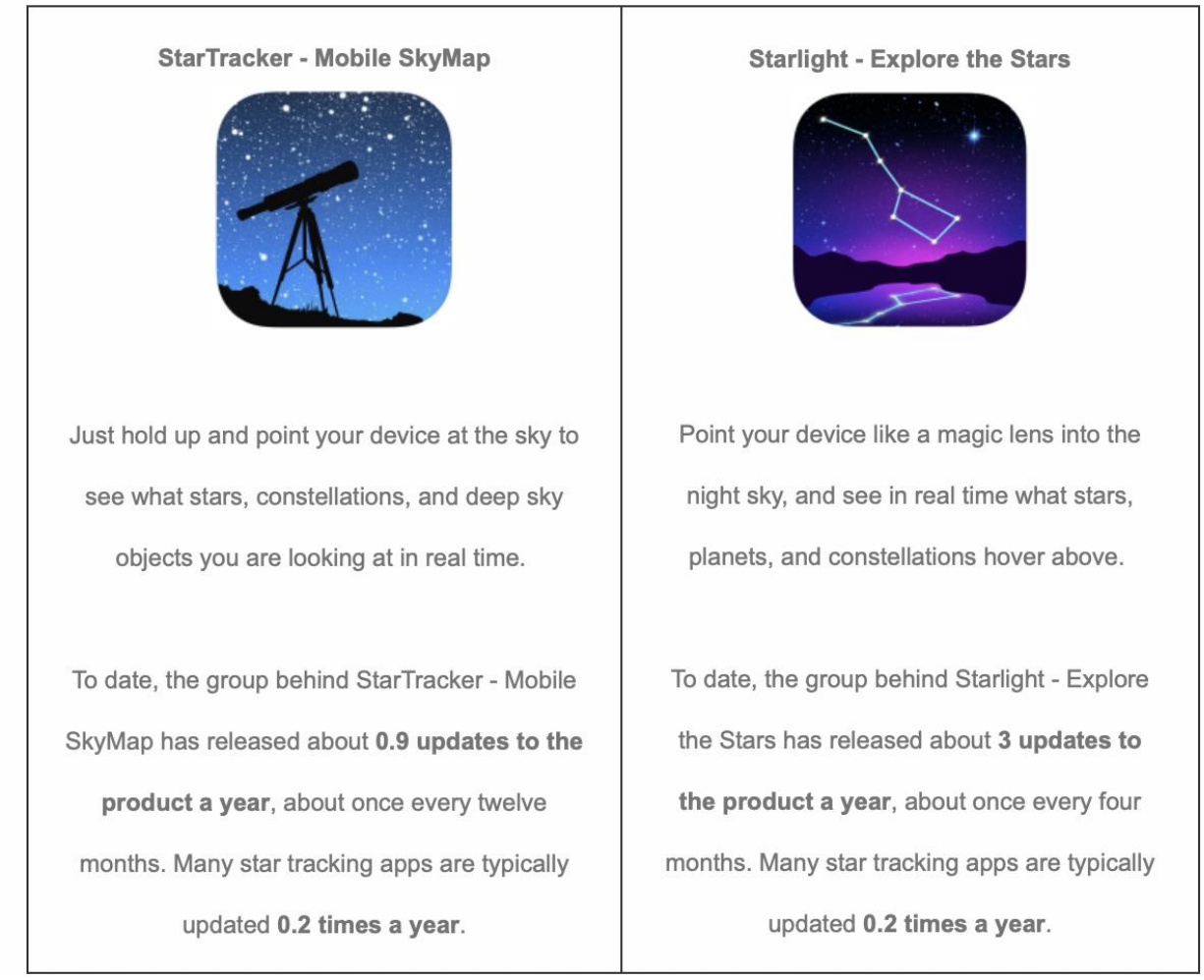


Figure S1. Descriptions of each app used in Study 4a. This example shows the category update frequency presented to those in low reference frequency condition. Those in the high reference frequency condition instead read that "Many star tracking apps are typically updated **4 times a year**."

Supplement 3: Calibration test results

Method

In this study, we test impressions of the two different categories of apps used in study 4b and its conceptual replications. There was no information about updates included in descriptions of the apps

Participants. Participants were recruited via Amazon Mechanical Turk to complete a brief study for $0.50. In all studies, incomplete responses, those with duplicate IP addresses, those with duplicate response IDs, and those who failed any attention checks were excluded from analyses, as pre-registered. We collected 94 complete responses (M_{age} = 37.35, SD_{age} = 12.54; 50% female).

Procedure. The calibration test used a 2 factor (fitness app pairs) between subjects design. Participants saw two pairs of real apps from the Apple app store and a short description of each app (one pair of stargazing apps, one pair of fitness apps, see Figures S2), selected a preferred option between each set, and made judgments about each pair of apps. All participants saw the same pair of stargazing apps. Participants were randomly assigned to see one of two pairs of fitness apps. The order of categories and apps in each pair were counterbalanced.

After seeing a category of apps, participants were asked which app they would prefer to download. Then, they answered five questions: How confident are you that you would prefer to download your choice? (on a 5 point scale from “totally unsure” to “totally confident”), How similar do you find these two apps (on a 5 point scale from “totally different” to “exactly the same”), How familiar are you with the topic these apps cover (on a 5 point scale from “totally unfamiliar” to “totally familiar”), How often would you use this type of app (on a 5 point scale

from “very infrequently” to “very frequently”), and How likely would you become bored of this type of app (on a 5 point scale from “very unlikely” to “very likely”).

Results.

Choice for any app in each pair did not significantly different from chance (Stargazing: $\chi^2 = 0$, df = 1, p = 1; Fitness pair 1: $\chi^2 = 2.17$, df = 1, p = .14; Fitness pair 2: $\chi^2 = 0.33$, df = 1, p = .56; proportions choosing each app in Table S1).

T-tests for ratings of app categories are shown in Table S2. Substantively, confidence in preference of app was directionally greater for either pair of fitness apps than for the pair of stargazing apps. There was no difference in choice confidence between pairs of fitness apps. People found the stargazing apps significantly more similar to each other than the apps in either pair of fitness apps. There was no difference in similarity between apps when comparing the two pairs of fitness apps. People reported they were significantly more familiar with the topic covered by either pair of fitness apps than with the pair of stargazing apps. There was no difference in familiarity when comparing the two pairs of fitness apps. People reported they would use apps like either pair of fitness apps significantly more frequently than apps like the pair of stargazing apps. There was no difference in anticipated frequency of use between the two pairs of fitness apps. People reported they would become similarly bored with all pairs of apps. Directionally, people found fitness app pair 1most likely to bore them on average, followed by the pair of stargazing apps, and the pair of fitness app pair 2.

Findings from this calibration test suggest that individuals find apps in the fitness category more distinct from each other than apps in the stargazing category, report they would use apps in the fitness category more frequently than apps in the stargazing category, and were more familiar with the content of fitness than stargazing.

	Selected less updated (unlabled)	Selected more updated (unlabled)
Stargazing pair	50%	50%
Fitness, pair 1	61%	39%
Fitness, pair 2	46%	54%

Table S1. Proportions of participants choosing each app in each pair. The more updated options were: StarTracker, in the stargazing pair, Full Fitness in fitness pair 1, and Fitness Buddy in fitness pair 2.

	Stargazing app pair vs. Fitness app pair 1	Stargazing app pair vs. Fitness app pair 2	Fitness app pair 1 vs. Fitness app pair 2
How confident are you that you would prefer to download your choice? (5, Totally confident, to 1, Totally unsure)	t(94) = 1.20, p = .234, Cohen's d = .198 Stargazing app pair M=3.66 ; Fitness app pair 1 M=3.83	t(94) = 1.50, p = .138, Cohen's d = .258 Stargazing app pair M=3.66 ; Fitness app pair 2 M=3.89	t(94) = -0.36, p = .722, Cohen's d = 0.07 Fitness app pair 1 M=3.83 ; Fitness app pair 2 M=3.89
How similar or different do you find these two apps? (5, Exactly the same, to 1, Totally different)	t(94) = -3.88, p < .001, Cohen's d = 0.72 Stargazing app pair M=4.07 ; Fitness app pair 1 M=3.50	t(94) = -3.04, p = .003, Cohen's d = 0.52 Stargazing app pair M=4.07 ; Fitness app pair 2 M=3.70	t(94) = -1.23, p = .22, Cohen's d = 0.25 Fitness app pair 1 M=3.50 ; Fitness app pair 2 M=3.70
How familiar are you with the topic these apps cover? (5, Totally familiar, to 1, Totally unfamiliar)	t(94) = 4.45, p < .0001, Cohen's d = 0.77 Stargazing app pair M=2.87 ; Fitness app pair 1 M= 3.71	t(94) = 4.72, p < .0001, Cohen's d = 0.80 Stargazing app pair M=2.87 ; Fitness app pair 2 M=3.72	t(94) = -0.04, p = .964, Cohen's d = .01 Fitness app pair 1 M=3.71 ; Fitness app pair 2 M=3.72
How often would you use this type of app? (5, Very frequently, to 1, Very infrequently)	t(94) = 3.20, p = .002, Cohen's d = 0.57 Stargazing app pair M=2.72 ; Fitness app pair 1 M=3.38	t(94) = 2.39, p = .018, Cohen's d = 0.42 Stargazing app pair M=2.72 ; Fitness app pair 2 M=3.20	t(94) = 0.79, p = .433, Cohen's d = .16 Fitness app pair 1 M=3.38 ; Fitness app pair 2 M=3.20
How likely would you become bored with this type of app? (5, Very likely, to 1, Very unlikely)	t(94) = 1.13, p = .262, Cohen's d = 0.20 Stargazing app pair M=2.83 ; Fitness app pair 1 M=3.06	t(94) = -0.13, p = .899, Cohen's d = 0.02 Stargazing app pair M=2.83 ; Fitness app pair 2 M=2.80	t(94) = 1.10, p = .276, Cohen's d = 0.23 Fitness app pair 1 M=3.06 ; Fitness app pair 2 M=2.80

Table S2. T-tests and Cohen's-d values for comparisons between app types.

Consider the following products:

Home Workout PRO	Full Fitness
	
Home Workout PRO includes video instructions for over 100 exercises for the whole body. Use workouts created by experienced fitness instructors or create your own personalized workouts.	Full Fitness provides video instructions for hundreds of exercises sorted by body region and target muscle. Use routines pre-defined by licensed fitness professionals or create your own custom exercise routines.

Figure S2a. Fitness app pair 1.

Consider the following products:

Fitness Buddy: Gym Workout Log	Streaks Workout
	
Be stronger. Be leaner. Be the best you. Build your own workout routines. Find out why gym goers are switching over to Fitness Buddy!	The personal trainer that you actually want to use. Exercise anywhere! Customize which exercises you want to do, making it great for all ages and abilities.

Figure S2b. Fitness app pair 2.

Consider the following products:

Starlight: Explore the Stars	StarTracker - Mobile SkyMap
	
Point your device like a magic lens into the night sky and see in real time what stars, planets, and constellations hover above. Stargazing has never been so easy!	Just hold up and point your device at the sky to see what stars, constellations, and deep sky objects you are looking at in real time.

Figure S2c. Stargazing app pair 1.

Supplement 4: Study 4b additional results

Minimum residual factor analysis was conducted for the five questions relating to product quality and effort to produce products, separately for each app type. This yielded two factors explaining 71% of the variance for the set of five variables for stargazing apps and 69% of variance for the set of five variables for fitness apps. Factor 1, with high loadings on difference scores of overall perceived quality (Stargazing: 0.87; Fitness: 0.87), how the product compared to its peers (Stargazing: 0.91; Fitness: 0.92), and how up-to-date the product seemed (Stargazing: 0.65; Fitness: 0.47), was labeled perceived quality. Factor 1 explained 46% of the variance for stargazing apps and 44% of the variance for fitness apps. Factor 2, with high loadings on difference scores of perceived investment from the producer (Stargazing: 1.00; Fitness: 1.00) and perceived overall effort (Stargazing: 0.35; Fitness: 0.22), was labeled perceived effort. Factor 2 explained 25% of the variance for stargazing apps and 25% of the variance for fitness apps. Factor 1, perceived quality, and Factor 2, perceived effort, were highly positively correlated (Stargazing: $r = 0.67$; Fitness: $r = 0.61$).

Supplement 5: Study 4b Conceptual replications

Method

Two replications of study 4b were pre-registered and conducted: (1) using different apps from study 4b, without bolded frequencies, and a $0.10 screener payment, #29101 https://aspredicted.org/blind.php?x=9e2sg4, and (2) using the same apps as in study 4b, without bolded updated frequencies, #30492 https://aspredicted.org/blind.php?x=9qk9pg.

Participants. Via Amazon Mechanical Turk, we collected 353 complete responses (M_{age} = 35.64, SD_{age} = 10.98; 61% female) in replication (R1) and 175 complete responses (M_{age} = 34.95, SD_{age} = 11.03; 46% female) in replication 2 (R2). Exclusions were the same as those used in study 4b.

Procedure. Both replications of 4b used the same choice procedure as in Study 4b. In the replications, only information about the update frequency of each app was underlined. Information about the category update frequency was not underlined. In R1, the fitness apps used were “Full Fitness” and “Sezzy Timer ” and the star apps used were “StarMap 3D Pro” and “Sky Guide” (details shown in Figure S3). The apps used in the second R2 were the same as those used in study 4b.

Figure S3 Apps used in CR1

Consider the following products:

Sezzy Timer:	**Full Fitness**:
	
The Sezzy Timer app shows where you are in your workout and also provides auditory cues to tell you what's coming up next. To date, the group behind Sezzy Timer has released about 2 updates to the product a year, about once every six months. Many fitness apps are typically updated 10 times a year.	With Full Fitness, hundreds of exercises are explained with clear pictures, videos, and text instructions. To date, the group behind Full Fitness has released about 1.5 updates to the product a year, about once every eight months. Many fitness apps are typically updated 10 times a year.

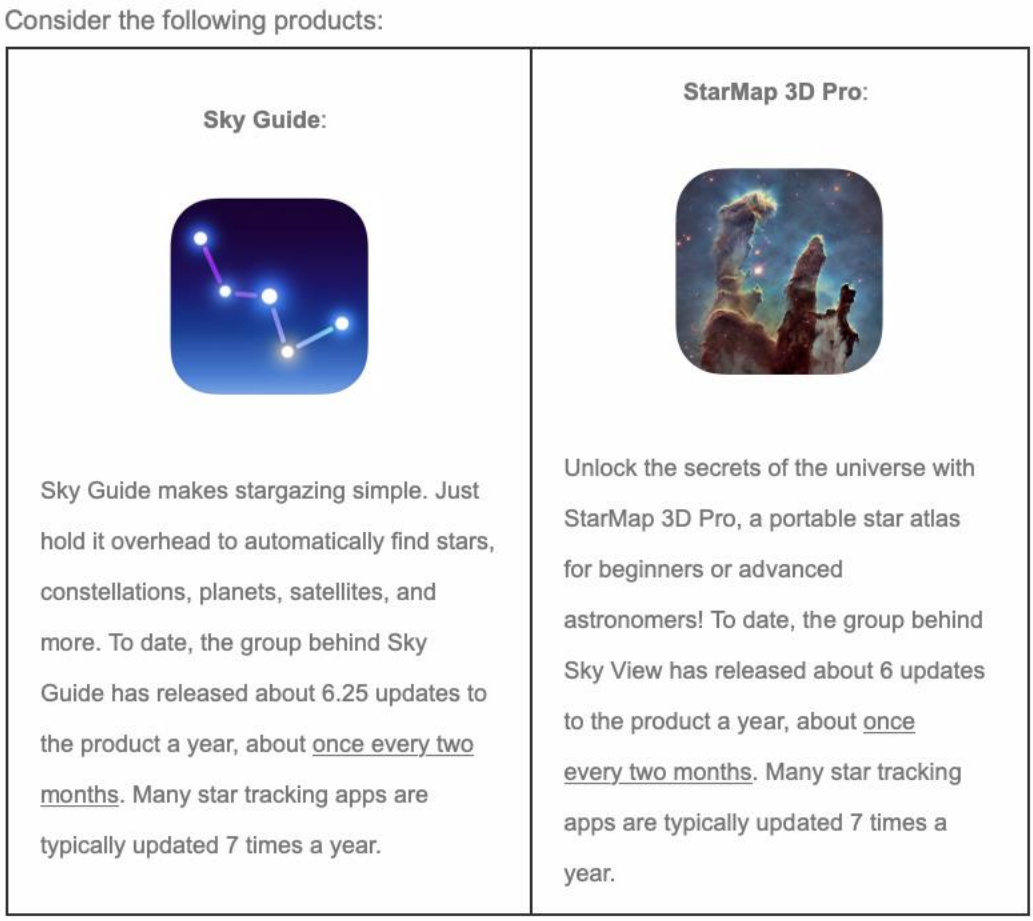


Figure S3. The app names, images, and app update frequency were true to tbe real products. The category updates for fitness apps was assigned to be either 1.25 or 10 per year. Category updates for stargazing apps was assigned to be either 5.23 or 7 per year.

Results.

In R1, participants were directionally more likely to choose the more frequently updated stargazing app (77%) in the high-reference-level condition, when both apps' update frequencies were below the reference level, than in the low-reference-level condition (75%; $\chi^2 = 0.21$, df=1, p = .646), when both apps' update frequencies were below the reference level. A similar pattern was seen for fitness apps (high: 34%, low: 30%; $\chi^2 = 0.59$, df=1, p = .443).

In R2, participants were directionally more likely to choose the more frequently updated stargazing app (82%) in the high-reference-level condition, when both apps' update frequencies

were below the reference level, than in the low-reference-level condition (73%; χ^2 = 1.47, df =1, p = .226), when both apps' update frequencies were below the reference level. A similar pattern was seen for fitness apps (high: 72%, low: 65%; χ^2 = 0.65, df =1, p = .421). We did not conduct further analyses of potential covariates.

Supplement 5: CDR Behavioral Labs study results

This study was preregistered on aspredicted.org: https://aspredicted.org/LJB_NZU, #32933

Participants. Participants were recruited at CDR Behavioral Labs at the Museum of Science and Industry. Participants could only take the study once. Incomplete responses and responses from people under 18 years old were excluded, as pre-registered. We collected 190 complete responses (M_{age} = 39.81, SD_{age} = 16.03; 61% female).

Procedure. This study used a two factor within-subjects design. All participants were shown a pair of devices two times, interspersed among questions from other unrelated studies. The devices were the same as in study 2. Participants were randomly assigned to learn that the category frequency of updates was higher than both options (10/year) in one set of questions and lower than both options (2/year) in the other set of questions. Participants read about and selected a construction measurement device from each pair.

Results.

We did not find a significant difference in choice of the more frequently updated product based on whether the category frequency was higher or lower than both options (χ^2 = 0.20, df = 1, p = .648). Directionally, people were more interested in having the more updated product when the category frequency was higher (37%) rather than lower (33%).