Journal of Economic Perspectives—Volume 38, Number 4—Fall 2024—Pages 155–174

Behavioral Incentive Compatibility and Empirically Informed Welfare Analysis: An Introductory Guide

Alex Rees-Jones

Consider an economist seeking to compare two economic situations and assess which is better for society. To pursue this goal, she first specifies what "being better for society" means. Formally, this entails quantifying the overall good or social welfare attained given the different possible allocations that could arise. She then forecasts the patterns of behavior that each situation will generate. By evaluating social welfare at the allocations that follow from these forecasted patterns of behavior, this economist now has what she needs to compare the social value of one option versus the other.

The approach just described succinctly captures economists' dominant paradigm for welfare analysis. Stated at this level of generality, it in some ways appears simple and straightforward: one just needs to specify how welfare will be defined and measured and then forecast individual behavior. Of course, forecasting the behavior of humans is challenging. At the same time, an enormous amount of economic research has been conducted with the explicit purpose of informing this stage of the modeling process, providing extensive foundations from which to build.

When generating the needed forecasts of individual behavior, the favored approach in economics is to assume that behavior will satisfy *incentive compatibility*. Put most simply, this means that researchers assume that individuals behave in the

■ *Alex Rees-Jones is an Associate Professor of Business Economics and Public Policy, Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania. He is also a Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. His email address is alre@wharton.upenn.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.38.4.155.

manner that best pursues their own interests. This approach is favored for good reason: it is extremely powerful. The assumption that individuals choose optimally immediately makes available all of the standard economic tools of "revealed preference." These tools provide well-developed means of estimating models of individual welfare, and those models can be used to determine what individuals will then choose. This framework thus provides the needed forecasts of the behavior that will be chosen in different situations.

Despite the power of this approach, reliance on incentive compatibility has a clear lack of appeal in certain settings. For many economic questions, individuals' optimization failures are central to the debate and thus cannot be ignored. For other economic questions, individuals' optimization failures might not be central to the existing debate, but incidental misoptimization may change the conclusion of that debate. In such cases, reducing our reliance on incentive compatibility may help us better analyze the economic environment and guide us towards better policies.

Motivated by these considerations, researchers increasingly conduct analysis that may be characterized as relying on *behavioral incentive compatibility*. Under this approach, behavior is forecasted not by assuming that individuals maximize welfare. Instead, the researcher attempts to model both the individuals' welfare and also the forces that guide them towards unwise decisions. These forces at times include psychological factors, incorrect beliefs about aspects of the decision problem, or preferences for things that are judged as normatively irrelevant and thus excluded from a standard welfare consideration. Despite the addition of these factors, the spirit of this exercise is extremely close to that driven by standard incentive compatibility. Just as in the standard case, this approach is based on assuming that individuals' decisions are compatible with their pursuit of incentives. This approach merely embeds some imperfection in their means of that pursuit, often drawing from work in behavioral economics.

In this article, I aim to introduce readers to empirically-informed welfare analysis based on behavioral incentive compatibility and to provide guidance for how to pursue a project involving such analysis. My interest in doing this comes from my experience having written several papers of this variety, actively engaged with this literature through most of its recent rise in prominence, and advised a number of students in their pursuit of this style of project. Having watched the literature evolve through that lens, two things stand out to me.

First, the potential value of this approach no longer needs to be taken on faith, but instead can be inferred from existing literature. Projects are being executed that address important economic questions, do so up to high standards of rigor, and ultimately have influence in diverse literatures. At least in some fields, I believe the approach has demonstrably grown beyond being "something popular with behavioral economists" and into something used, when appropriate, by standard members of the field.

Second, despite that success, there is an unfortunate hurdle that I believe has persistently slowed progress. Different fields have different core behavioral concerns, playing out in potentially very different economic environments. This naturally

contributes to a sense that solutions and approaches must be context-specific. Behavioral economics is often criticized for providing too many ad hoc theories instead of a unified framework that can immediately be brought to new settings; I believe this contributes to a common sensation that welfare analysis informed by behavioral economics would also be ad hoc. Yet, looking across successful examples of this type of research, it appears that the common practices for pursuing these projects are ultimately very similar across the subfields that have adopted them, and that there is an underappreciated degree of commonality in the template that is followed. I will seek to make the main elements of it clear, in the hopes of helping to make the pursuit of these projects less daunting.

Three Examples of Welfare Analysis with Behavioral Incentive Compatibility

To begin, I present three examples of projects where welfare assessments hinge critically on applications of behavioral incentive compatibility. These projects make concrete some of the issues just discussed—for example, they illustrate some types of analysis that can benefit from this approach; they illustrate how application of the approach can significantly change the conclusions we reach; and they illustrate that this approach has found traction across a range of fields with very different methods, settings, and interests.

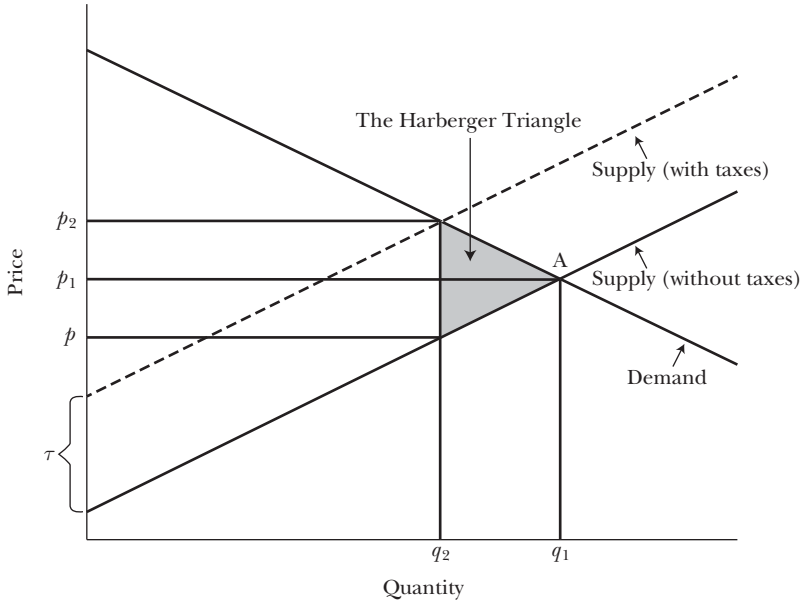
Behavioral Incentive Compatibility and Sales Taxes

We first review the pioneering work of Chetty, Looney, and Kroft (2009), in which the authors consider a classic question: How to calculate the welfare losses from sales taxes?

A common approach to answering this question is to use the "Harberger triangle" approximation (Harberger 1964), which fundamentally relies on an appeal to incentive compatibility. To illustrate, consider a standard supply and demand framework as represented in Figure 1. If we assume that all purchase decisions are incentive compatible, the demand curve serves two important functions in welfare analysis.[1] First, the demand curve provides a direct measure of consumer welfare. The difference between the willingness to pay encoded in the demand curve and the amount actually paid (that is, "consumer surplus") is a natural money-metric measure of the consumer benefits arising from the trade. Second, the demand curve allows us to infer what purchases will be made in counterfactual situations, such as when considering a new tax to be imposed in a previously untaxed market. Figure 1 illustrates the case where the new tax, τ, is imposed on the supply side and thus shifts the supply curve upward. This raises the equilibrium tax-inclusive price

[1] The assumption that all sales decisions are rational serves an analogous role for forecasting welfare and behavior of the supply side. I focus on the demand side here as it is the focus of Chetty, Looney, and Kroft (2009).

Figure 1
The Harberger Triangle



Source: Reproduced from Hines (1999).
Note: This figure presents a standard demonstration of Harberger triangle analysis. In this demonstration, a tax of size τ is introduced on the supply side of the market. The Harberger triangle is represented in the shaded region and captures the lost surplus from the trades that were eliminated by the post-tax price increase.

from p_1 to p_2, which rationally dissuades consumers who have willingness to pay between p_1 and p_2 from purchasing the good. The consumer surplus lost by these dissuaded consumers, along with the producer surplus lost by the producers who no longer trade with those consumers, is the "excess burden" or welfare loss from the imposition of this tax. It is represented in the shaded triangle in Figure 1.

Harberger-style analysis has been used extensively in economics and is clearly valuable. However, Chetty, Looney, and Kroft consider a specific reason why this analysis might be incomplete and why there might be room for improvement: consumers may not optimally react to taxes when the taxes are not salient. Consider, for example, a consumer who selects which groceries to purchase based on their price tags. In many parts of the world, these price tags would report the amount of money that must be paid to take ownership of the goods. In a US store, by contrast, price tags typically exclude sales taxes, which are then later imposed at the register. This labelling can naturally be expected to lead to mistakes if some consumers do not know sales tax rates, or do not know that some groceries are taxable and others are not, or know that there are taxable and untaxable groceries but do not know which are which, or know all of this but forget to attend to it, or remember

to attend to all of this but make mistakes in calculations, or do some rounding along the way, or correctly process everything but only notice changes in taxes slowly, or are able to correctly process everything but deem doing so not worth their time, and so on.

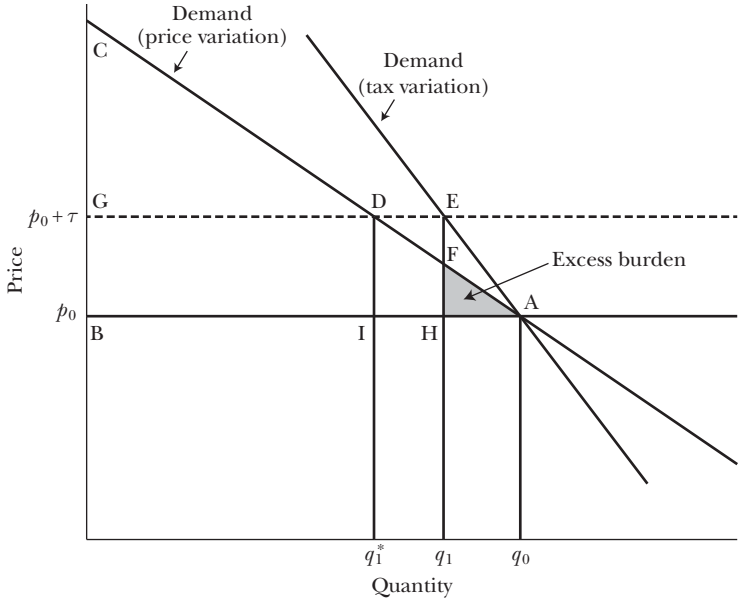
If consumers are failing to attend correctly to sales taxes collected at the register, what are the consequences for behavior? A natural consequence would be insensitivity, or lack of elasticity, to the sales tax amount, even in cases where the consumer would be sensitive to exactly comparable changes in the price advertised on the price tag. Chetty, Looney, and Kroft provide two empirical exercises—one field experiment and one observational study—that each directly demonstrate underreaction to taxes and allow for estimation of a parameter that governs the resulting reduction of elasticity.

Bringing these findings and observations together, Chetty, Looney, and Kroft demonstrate how nonsalient taxes can be accommodated in Harberger triangle calculations. Their modification may be understood as replacing the prior reliance on incentive compatibility with reliance on behavioral incentive compatibility. Formally, they model individuals as still making nearly rational decisions by assessing if the value of a good exceeds its price. However, they assume that only a portion of the tax is accounted for when price is calculated, rendering the overall decision rational except for a price misperception. The resulting demand can be estimated by examining purchase decisions as taxes vary, and may be used to forecast how demand behavior will change as taxes are changed. However, unlike in the standard case, this demand curve no longer is assumed to reveal welfare, because it is influenced by mistakes. A demand curve that is not influenced by mistakes can be estimated by examining how demand responds to variation in posted prices, which Chetty, Looney, and Kroft assume are processed correctly.

Figure 2 summarizes this analysis. Consider a market that would be at point A in the absence of a tax. From this point, Chetty, Looney, and Kroft draw two demand curves. The steep one represents demand arising from (nonsalient) tax variation. The shallower demand curve represents demand arising from price variation. The difference between these demand curves reflects the empirical finding that the quantity demanded will change more in response to a change in salient price than a comparable change in nonsalient tax. With these two curves graphed, Chetty, Looney, and Kroft then assess surplus or welfare by making use of the demand curve arising from price variation while still assessing predicted behavioral changes from the demand curve arising from tax variation.

To illustrate the calculation of the Harberger triangle, imagine a small tax τ were imposed on the economy in equilibrium A. Assume this economy faces a flat supply curve, as Chetty, Looney, and Kroft assume to focus attention on the demand side. In this case, if consumers responded to the new tax using the welfare-relevant demand curve (as arises from price variation), the new equilibrium would occur at point D. The standard Harberger triangle would be AID. The key observation of Chetty, Looney, and Kroft is that this calculation would overestimate the demand response of consumers by not accounting for their propensity to underreact to the

Figure 2
A Harberger Triangle When Taxes Are Non-Salient



Source: Figure 4 from Chetty, Looney, and Kroft (2009) (with modifications to labeling).
Note: This figure presents analysis of the consequences of introducing a tax of size τ to a market in which consumers react to tax salience. In this analysis, consumers are assumed to respond optimally to variation in posted prices, and thus the demand relationship from price variation can be used to infer welfare. Consumers are not assumed to respond optimally to variation in nonsalient taxes, leading a separate demand relationship to arise from tax variation.

nonsalient tax. Taking that into account would require using the other demand curve, which suggests equilibrium would occur at point E. If the demand curve with tax variation were believed to be welfare-relevant, this would suggest that the Harberger triangle should be AHE. However, the demand curve with tax variation is not welfare-relevant, and the demand curve for price variation must be used for the calculation of surplus. As a result, triangle AHF provides the desired estimate of excess burden: it uses the demand curve from tax variation to determine the quantity demanded in equilibrium q_1, but assesses lost consumer surplus by integrating the price-based demand curve between that new quantity demanded and the initial quantity demanded q_0.

In applying these results, this analysis yields a perhaps surprising conclusion. Despite the common intuition that failure to optimize is harmful, this analysis shows that welfare losses stemming from taxes are *reduced* when the taxes are nonsalient.

This arises because some surplus-reducing but individually-rational decisions to quit purchasing the good are not made.[2]

Behavioral Incentive Compatibility and Health Insurance

A second application of behavioral incentive compatibility for welfare analysis can be found in the work of Handel (2013). In this paper, Handel assesses a classic topic in health insurance markets: how they operate in the presence of adverse selection.

To illustrate the issue of adverse selection, consider a population of risk-neutral individuals buying a health insurance plan. If purchased, this plan will cover all healthcare costs. The operator of the health insurance plan knows the average cost of providing healthcare and offers a plan at that cost. Further assume that individuals decide whether to purchase the plan in an incentive-compatible way: they buy the plan if their expected costs of healthcare are higher than the cost. This understandable behavior leads to an unfortunate market dynamic. The insurer will soon find that, while the plan was priced appropriately for the average person in the population, the plan is not priced appropriately for the individuals who purchased the plan. These customers have been selected for the adverse trait of having higher-than-average health costs. The insurer must raise prices to prevent operating at a loss. This leads to further selection by "pricing out" even more customers, perhaps making further price increases necessary. Repeated rounds of this repricing can lead a large fraction of the populace to rationally remain uninsured due to the unavailability of an acceptably priced insurance product—a phenomenon often referred to as a "death spiral."

Adverse selection is a phenomenon that is driven by an unfortunate pattern of incentive-compatible behavior. Handel is partially motivated by the observation that, for health insurance choice, idealized incentive compatibility might fail. To illustrate again with an example, consider a worker trying to pick the best health plan of the five offered by her employer. These plans often have differences in deductibles, copays, coinsurance rates, coverage, and more. Thus, at a minimum, determining the optimal choice requires some understanding of how these provisions operate and interact. Furthermore, the consequences of these various provisions must be assessed across a large number of health situations this employee might face. The employee must assess her optimal plan if she is healthy all year, her optimal plan if she develops a specific rare illness, her optimal plan in a large number of other health contingencies, and the likelihoods of all of these different contingencies. Given these challenges, one could imagine that this worker might avoid making a serious attempt to determine the optimal plan due to its perceived futility, or that

[2]Follow-up papers have demonstrated that this simple conclusion might not hold when the method of welfare analysis is enriched in certain ways (for example, see Goldin and Homonoff 2013; Reck 2016; Taubinsky and Rees-Jones 2018). However, even these follow-up papers adopt the behavioral incentive compatibility approach when assessing welfare and merely debate the specification of some model components.

she might incorrectly select the optimal plan even if she tries to determine it. In such cases, this worker will of course be individually worse off. But how does the aggregation of these mistakes affect the market as a whole, and adverse selection itself?

To study these questions, Handel studies imperfect choice of health care plans in a large US firm, where workers must make decisions somewhat like the example just considered. In his data, he directly documents a low propensity to change plans across time and provides some suggestion that this market is insufficiently active relative to a rational benchmark. But much more strikingly, he studies a case where a plan became formally dominated—that is, another plan became preferable to this plan no matter the health contingency that would arise. Remaining in the dominated plan is not incentive compatible, and yet many employees failed to abandon it. These features together provide a compelling demonstration of some degree of consumer inertia.

To model inertia, Handel augments an otherwise-rational model of insurance demand to include an "as-if" switching cost. In standard models in this environment, individuals will switch to a new insurance plan if the plan offers infinitesimally better terms than its best competitor. In Handel's model, consumers act as if they will only switch from their plan when the returns to doing so are sufficiently large. Handel's estimated model suggests that benefits of switching plans must be valued above approximately $2,000 to motivate a switch. Of course, switching plans does entail some time and effort, so some degree of switching cost can be rationalized. But it is hard to rationalize a switching cost that is so large. This supports treating an individual's reliance on this switching cost as a mistake, and supports the treatment of the switching cost as an element of estimated utility that should be excluded from welfare. Use of the model in this way serves as the centerpiece of Handel's application of behavioral incentive compatibility.

Handel uses this estimated model to assess the welfare effects of consumer misoptimization in this market. To do so, he evaluates the effect of reducing inertia by scaling down switching costs. As a baseline analysis, Handel considers this change while holding plan pricing fixed (and thus preventing the consequences of adverse selection from playing out). In this analysis, reducing inertia leads to improved sorting of individuals to their individually rational policies, resulting in a substantial improvement to consumer welfare. This accords with common intuitions that helping individuals avoid mistakes helps their welfare, all else equal, perhaps suggesting that "nudges" to combat inertia would be useful.

However, a quite different conclusion arises once the impacts of adverse selection are reintroduced to the model. When plans are allowed to endogenously reprice their products as consumer demand changes, Handel finds that reducing inertia exacerbates adverse selection. As individuals sort to new plans once inertia is reduced, some plans are effectively removed from the market due to losing their lower-cost customers who previously stayed in the plan due to inertia. As such individuals are lost, prices rise, leading to further re-sorting. The end result is a microcosm of a death spiral that drives substantial declines in overall welfare.

A simple takeaway from this paper—only assessable through application of behavioral incentive compatibility—is that inertia and consumer misoptimization may at times play an important role in keeping health insurance markets functional in the presence of potentially debilitating adverse selection. In markets with adverse selection, "the problem" is generally that individuals make their optimal choices based on private information (in this case about health costs). If a behavioral force like inertia prevents them from doing so, this can at times be helpful for overall welfare, even if the behavioral forces come with welfare losses of their own.

Behavioral Incentive Compatibility in School-Choice Market Design

A third example of behavioral incentive compatibility in welfare analysis can be found in the work of Kapor, Neilson, and Zimmerman (2020). This paper assesses some much-studied questions in market design: how should we assign students to schools, and should we favor the "immediate acceptance" or the "deferred acceptance" algorithm?

When determining students' assignments to schools, a large and growing number of school systems use a formal centralized matching system. In such a system, both students and schools are asked to submit their preferences for assignments. For a student, this could be an indication of their favorite school to attend, their second favorite school, and so on until their last acceptable school. For a school, this could be an indication of their favorite student to admit, their second favorite, and so on until their last acceptable student. Schools additionally report how many seats they have available. Once this information is submitted, the school district can use it to determine a desirable way to assign students to schools.

Incentive compatibility plays a crucial role in assessing these procedures. Typical analysis assumes that students rank schools while rationally responding to any strategic incentives introduced by the procedure. This practice is clearly important because students can often face strong incentives *not* to report their true preferences.

To illustrate this potential for incentives to misreport preferences, imagine that assignments are determined by the following procedure. First, the school district tries to assign each student to the school she ranked first. If the school said that the student is unacceptable, or if the school is already filled to capacity with other applicants that the school prefers, the student is not assigned a seat. Otherwise, the student gets a seat at the school. Those assignments are treated as final and each schools' capacity is updated to reflect the seats that have been removed from the market. In the next step, the school district repeats this procedure, now trying to match students who did not match to their first-choice school to the remaining seats at their second-choice school. The procedure continues iterating in this way, moving down the students' preference lists, until all students are matched or every student has attempted to match at every school that they ranked.

The procedure just described is famous within the school-choice literature for producing unfortunate incentives, and is called the immediate acceptance mechanism or the Boston mechanism. To illustrate the incentive problems, consider two

schools, A and B. Both are very popular: they can fill all of their seats with students who ranked them first. Now consider a student who prefers A to B. Fortunately, school B ranks this student very highly. Unfortunately, school A does not. In such a case, this student could be matched to school B if she ranks it first. However, if she ranks school A first and school B second, she will not match to school A in the first stage of the procedure and will have no remaining seats available at school B in the second stage. This student thus faces clear incentives not to list school A: doing so would cost her the chance to study at school B. Generalizing beyond this simple example, this procedure offers strong incentives not to rank options where a match is unlikely, and generally punishes sincere participants to the benefit of the strategically savvy (Pathak and Sönmez 2008).

Avoiding this incentive problem is one of several reasons why economists have favored the use of the deferred acceptance mechanism of Gale and Shapley (1962). This mechanism may be understood as a modification to immediate acceptance that does not remove filled seats after each round, but instead allows more-preferred new applicants to displace previous matches. This eliminates the problem discussed in the example above, where a desired applicant is only considered after seats have been irrevocably claimed by students who ranked the school higher. Under deferred acceptance, these claims are no longer irrevocable. This mechanism structure results in deferred acceptance being strategy-proof: regardless of the behavior of other market participants, students can do no better than truthfully reporting their preferences (Dubins and Freedman 1981; Roth 1982). For this reason and others, deferred acceptance has largely served as the tool of choice for school-choice market designers in recent decades. (For much fuller detail on the use of these mechanisms for school assignment, a useful starting point is Abdulkadiroğlu and Sönmez 2003.)

The contrast between these two mechanisms may suggest that the choice between them is obvious: use of deferred acceptance, where students can report their preferences truthfully, seems wise compared to use of immediate acceptance, where strategic behavior is necessary and sincerity is punished. One provocative counterpoint to this comparison comes from Abdulkadiroğlu, Che, and Yasuda (2011), who note that immediate acceptance can, under some conditions, extract cardinal preference information that can lead to a higher-welfare match. This could potentially lead a market designer to prefer immediate acceptance despite its incentive properties.

To illustrate the issue, consider two students vying for two positions at schools, again labeled A and B. Say the two students both rank position A over position B. Despite that symmetry in rankings, there can be significant asymmetry in the welfare consequences of assignments. For example, if one student has essentially the same welfare at each school, whereas the other student is vastly better off at school A than school B, there could be strong welfare motives for saving the seat at A for the student who benefits from it more. The operation of deferred acceptance has no feature that pushes for this outcome. By contrast, the optimal reporting strategy for immediate acceptance is a function of cardinal utility differences and can at times lead to welfare gains by guiding assignments with that information.

The discussion up until now explains the state of the literature at the time Kapor, Neilson, and Zimmerman entered. To summarize, in this literature, deferred acceptance was broadly preferred to immediate acceptance as a means of matching students to schools. However, some theoretical considerations suggested that immediate acceptance might have welfare benefits. The models that lead to these conclusions rely on students optimally strategizing about their preference submission, taking into account their probabilities of matching to different schools. But what if students and their families don't know these probabilities, or have systematically biased beliefs? This motivates Kapor, Neilson, and Zimmerman's central question: are the theoretical benefits of immediate acceptance "worth it" when failures of probability estimates are taken into account?

Kapor, Neilson, and Zimmerman address this question using data from the New Haven Public School System. During the window of study, New Haven based school assignments on a procedure that was essentially identical to immediate acceptance. Kapor, Neilson, and Zimmerman secured access to administrative data, thus giving them access to the reported preferences that are used by the algorithm to determine the match. Such data are extremely valuable for the pursuit of a standard study of a school choice mechanism. Despite being valuable, they are insufficient for Kapor, Neilson, and Zimmerman's purposes, because they do not directly reveal the (possibly incorrect) beliefs about admissions probabilities that families hold. To address this data need, they also fielded a survey among participants in this match. While the survey served several purposes, its key function was to elicit families' beliefs about admissions probabilities with different possible preference submissions. They use these data to document substantial inaccuracy in families' probabilistic beliefs.

These findings illustrate a potential need to import a behavioral incentive compatibility notion into welfare inferences for this setting. To estimate preferences and assess welfare in a setting like this, the current standard approach is to assume that the preferences that were submitted maximize expected utility (as in Agarwal and Somaini 2018). This provides revealed-preference valuations of the different schools, which may be used to measure the welfare of a given assignment. Kapor, Neilson, and Zimmerman instead assume that the rank-ordered lists that were submitted maximize expected utility *conditional on the model of incorrect perceptions of match probabilities.*

Assessing total welfare with both approaches, a striking pattern emerges. When relying on standard incentive compatibility, Kapor, Neilson, and Zimmerman find that immediate acceptance outperforms deferred acceptance. This, viewed in isolation, would be a provocative finding: the widespread preference for deferred acceptance on the grounds of its avoidance of strategic incentives might be reducing welfare. This provocative finding is immediately reversed when considering the analysis based on behavioral incentive compatibility: once analysis accounts for families' difficulty in assessing admissions probabilities, the benefits of immediate acceptance decline. Deferred acceptance then preserves its status as the favored mechanism. This serves as an example of a case where reliance on standard incentive compatibility might lead to an unwise policy decision, and one that would be avoided by taking into account additional behavioral considerations.

Guidance for Welfare Analysis Based on Behavioral Incentive Compatibility

The three examples just considered contain welfare analyses of quite different economic questions drawn from quite different economic fields. Despite the different foundations of each of these analyses, there are clear similarities in their manner of execution. While I have discussed only three examples, I believe this similarity to be reflective of a broader phenomenon. In my observation, successful welfare analyses using the behavioral incentive compatibility approach tend to draw upon a relatively small set of tricks and techniques to make this potentially very complicated exercise manageable. In this section, I aim to provide general guidance on the execution of this approach that makes these techniques clear. To do so, I walk step-by-step through the stages that a researcher must complete in order to execute this approach and draw attention to common solutions to the problems that arise at each stage.

Specifying the Model of Welfare

While this organization is not universally the case, many papers relegate their welfare analysis to a short, final section that is presented as a way of interpreting earlier estimates. As a means of efficient scientific communication, I believe this practice often makes sense. However, this structure of writing can lead one to infer that, during the research process, the development of welfare analysis begins after the empirics are largely completed. While this ordering sometimes works, I do not recommend it. These analyses normally involve a model that is comparatively complex. Empirics that are not tailored to the model's requirements will often fail to provide everything that is needed. What is worse, one may determine late in the process that some needed pieces cannot be generated even with modifications to one's empirics.

Given these concerns, I strongly recommend writing out one's desired model of welfare as early as possible in a project so that it might inform the design of the empirical strategy (which might itself then point to necessary changes to the model). In simplest terms, specifying this model will involve providing a precise means of evaluating the social welfare arising from a given allocation and a precise means of forecasting the allocation that will arise from individuals' behavior. After specifying both the welfare criterion and the behavioral model, the research can then turn to estimating the behavioral model.

To begin this process, the first step is specifying a welfare criterion; that is, one must specify how to assess if a situation is better or worse. In common economic applications, this is often done by summing the costs and benefits as in cost/benefit analysis, summing the surplus from trades as in supply/demand analysis, or by summing some measures of individuals' welfare as in utilitarian analysis.

When ranking alternatives using a welfare criterion, a researcher is codifying their moral values. Quite inconveniently for economists, not all humans share the same moral values, and concordantly not all researchers agree on what constitutes

good welfare analysis. Some might be happy to measure welfare with the sum of surpluses as in the Harberger triangle analysis, while others might balk at ignoring *who* gets the surplus (say, the rich or the poor?). Some might prefer to proceed with a sum of utility functions that reflect a declining marginal utility from wealth, while others might balk at the different treatments individuals get in such an approach. Disagreements like these, and many more, provide a large amount of material for debate on essentially any welfare analysis one could write.

The subjectivity inherent in welfare analysis means that deploying it can be contentious, but it need not always be so. In some subfields, or in some topic areas, the need for welfare analysis has been sufficiently strong that researchers have had to engage with it often. And in doing so, they have often developed strong norms on how such analyses should be conducted and have developed extensive literatures to support such decisions. If one wants to assess tax policy, for example, there are extremely well-developed frameworks available that the research community demonstrably will tolerate. If one wants to assess a topic that does not have an existing playbook for welfare analysis, tolerance is not guaranteed.

This leads to one important recommendation for the process of project development: assess early on whether the project requires *just* innovation in the way behavior is modeled or whether it also requires innovation on standard welfare analysis. One could proceed in either case, but it is important to be clear-eyed that simultaneously innovating on two fronts is substantially more difficult than "merely" innovating on one. Battles on multiple fronts should be initiated with great caution and only with a compelling need. This advice is supported when examining our leading examples. In each of these papers, the analysis was carefully designed to look "normal" to members of the relevant literatures if the isolated behavioral element were removed. In each case, I believe the wisdom of the paper might not have been as widely appreciated if this decision had not been made.

Specifying the Model of Behavior

With a welfare function in hand, we may now perform welfare comparisons as long as we know the inputs to the welfare function that arise in each studied situation. In traditional economic analysis, these inputs are often the allocation of goods, which is assumed to be influenced by the choices of individuals pursuing their rational incentives.

The defining characteristic of welfare analysis based on behavioral incentive compatibility is that allocations are assumed to be influenced by the choices of individuals pursuing their incentives while also being affected by behavioral economic forces. The boundaries of what constitutes "behavioral economic forces" are somewhat nebulous, but I personally interpret this very broadly. Clearly within the boundaries are issues that draw directly on cognitive or social psychology; issues related to biased or imperfect forecasting of probabilities or states; issues that relate to social preferences; issues that relate to nonexponential time discounting; and issues that relate to imperfect cognition, perception, or attention. In our three focal examples, some behavioral economic forces were:

(1) a tendency to underreact to nonsalient taxes, which could occur if individuals forget to attend to them, (2) a tendency to fail to change health insurance plans when it is financially advantageous to do so, which could occur if individuals fail to attend to their insurance or find doing so psychologically aversive, and (3) a tendency to incorrectly assess one's probability of acceptance at a school, which could arise from a wide variety of the failures of probabilistic reasoning or information frictions.

Because there are so many ways for decision making to be imperfect, there are an enormous number of possible models that could be deployed within the behavioral incentive compatibility approach. Despite the idiosyncrasy in models that this causes, there are some important regularities in how the models are developed. I highlight two regularities: (1) *using simple models relative to behavioral-economic norms* and (2) *making defensible normative judgements.*

Simple models relative to behavioral-economic norms. When studying imperfections in decision-making, there are often multiple possible underlying forces that could generate the behavior of interest. Modeling the full details of these competing forces can be critical in a study oriented towards best understanding the root cause of the phenomenon. Such a model can illustrate what is necessary to identify separately one force from another, and if estimated it could provide a comparatively detailed and accurate means of predicting behavior. But while there are clearly circumstances where a detailed and process-focused modeling approach is appropriate, proceeding in this way is rarely ideal for pursuing welfare analysis. Some distinctions that are extremely consequential in, say, a study of psychology are not consequential for welfare. In the common situation where tractability is a problem, a researcher faces strong incentives to remove such distinctions from at least the basic version of the model under study.

To illustrate, consider again the underreaction to sales taxes studied by Chetty, Looney, and Kroft (2009). As discussed earlier, there are many reasons why this underreaction could arise, and these reasons might be active at the same time. To repeat a few: perhaps some individuals do not know that the sales tax applies to the item considered, and perhaps some individuals decide not to take the moment to consider the sales tax because they deem it not worth their time, and perhaps some individuals wish to attend to sales taxes but persistently forget to do so. A model that fully incorporated the nuances of these different causes of the behavior would be challenging to identify empirically and would complicate theoretical analysis. However, Chetty, Looney, and Kroft argue that they do not need to model each of these distinctions fully, because the welfare-relevant consequence of any of these stories is a wedge between "true" price elasticity and the analogous elasticity in the presence of nonsalient taxes. Thus, Chetty, Looney, and Kroft work with a maximally simple model of this phenomenon: elasticity is scaled down by a single parameter when price variation is coming from a nonsalient tax. If Chetty, Looney, and Kroft's goal were to fully understand the determinants of this inelasticity, or to determine how to design interventions to combat it, this modeling decision would be limiting. But given that their goal was to incorporate the consequences of nonsalience into

Harberger triangle analysis, this simplification instead makes progress possible where it would not be otherwise.

This value of simplification is also on clear display in the other two example papers. Handel (2013) studies individuals failing to change their insurance plan when it is financially advantageous to do so. Many failures of decision making could lead to this behavior, and yet Handel restricts these forces to operate through a single "as-if" cost-of-change parameter. Kapor, Neilson, and Zimmerman (2020) study families applying to schools in an imperfect way due to their inaccurate assessments of their probability of admission. These inaccurate assessments of probabilities could arise for many reasons and may have many causes, and yet Kapor, Neilson, and Zimmerman work with a simple model and explicitly discuss some issues excluded for tractability. I believe the fact that all three of these papers work with simplified behavioral models reflects a broader regularity: researchers attempting empirical welfare analysis based on behavioral incentive compatibility face a challenging enough task that they often cannot proceed without some degree of model simplification.

The advice to work with a simple model that is tailored to welfare analysis may not feel useful to a researcher who currently has a complex model in hand. In such a situation, how can the complex model be improved? One systematic way to pursue this question is to attempt a sufficient statistics approach, in which the researcher considers the desired welfare analyses and determines, in those formulas, the minimal amount of information that needs to be measured. In some cases, one can find that not all model primitives need to be estimated—a common example is finding that a local elasticity is sufficient for analysis rather than needing to know the further parameters of a utility function. This approach has long been used to facilitate welfare analysis with standard, fully rational economic models. I believe the realization that this approach works quite well for behavioral economic models is one of the factors contributing to the recent surge of work applying behavioral incentive compatibility. For more guidance on the sufficient statistics approach, see Chetty (2009).

Defensible normative judgments. By assuming that individuals pursue behavioral incentives that are different than those encoded in the welfare function, the researcher is assuming that individuals pursue goals that should not be objectively valued. Modern economists have been wary of taking this type of paternalistic stance, and for good reason. Social planners acting on paternalistic motives have at times been mistaken, misguided, or evil, and this has generated a basis to view such analysis as dangerous. What's more, there is an off-putting hubris inherent in paternalistic policy analysis: who is the researcher to say, confidently, that they know what is best for others? These concerns are among the factors that have pushed economists to be so firmly wedded to revealed-preference approaches. And as a result of that training, most economists will only abandon the presumption of welfare-maximizing behavior after being confronted with a quite strong case.

This status quo means that a researcher must make a very strong case for her behavioral incentive compatibility assumptions. In the best-case scenario, this will involve (1) a strong conceptual case for why imperfect decision making might occur, (2) a strong rationale for why pursuit of this imperfection should not be weighted

by the social planner, and (3) a strong empirical demonstration that supports the conceptual case. All three of the running examples were written in accordance with this advice. They each consider a relatively simple decision error that seems natural to many readers. The behaviors they consider are relatively unambiguously "errors" that are difficult to attribute to unusual preferences. And each paper contains clear empirical "smoking gun" evidence that its hypothesized imperfection is active. I believe their ability to deliver on these three requirements was critical to the success of these papers, and these characteristics are common among similar successful cases.

Compared to these examples, researchers face a more challenging situation if they cannot compellingly demonstrate the presence of the hypothesized behavioral channel, or cannot compellingly resolve its welfare-relevance. However, even in those cases, possible paths forward are available.

When the behavioral channel is in doubt, the welfare exercise can still be pursued contingently: *if* individuals behave in this way, *then* these welfare results follow.[3] This path may be of limited interest if few readers accept the "if" clause, but at least it allows for communication of results to those that accept that clause.

When the welfare-relevance of the behavior is unclear, welfare analysis can often become quite challenging to pursue. This problem has plagued some of the most common models in behavioral economics. To illustrate, consider the phenomenon of loss aversion that is famously incorporated into prospect theory (Kahneman and Tversky 1979; for a review in this journal, see Barberis 2013). Loss aversion is modeled as a tendency for individuals to value marginal reductions of a loss discretely more than they value marginal increases of a gain, thus making the assessment of the same absolute amount differ depending on whether it is framed as a loss or a gain. Despite the very large amount of research on loss aversion, there remains active disagreement as to whether it reflects a welfare-relevant preference or a mistake in reasoning. This disagreement has been a hinderance to individuals who seek to conduct welfare analysis with prospect theory (including me). Encouragingly, recent papers have provided useful guidance on how to best proceed in the presence of such modeling uncertainty. The core idea of these papers is to parameterize welfare-relevance and consider a range of values for the relevant parameter. With this framework, one can characterize welfare under the assumption that the behavioral component is zero percent welfare-relevant, 100 percent welfare-relevant, and everything in between. Presenting results in this way allows a reader to assess the conclusions that align with their beliefs on welfare relevance and allows the researcher to clearly communicate when claims are sensitive or insensitive to these assumptions. For development of this approach, see Goldin and Reck (2022) or Reck and Seibold (2023).

[3] Of course, all welfare analysis is contingent on its behavioral assumptions, but it is common (and reasonable) to emphasize this contingency to different degrees depending on the degree of confidence in those assumptions.

Estimating the Model for Welfare Analysis

By completing the steps in the previous sections, a researcher has laid out the key objects necessary to conduct a welfare analysis. We now turn to the question: how might these objects be estimated? As before, the great variety of settings and behaviors that could be modeled preclude a complete answer to this question. However, again, there are clear commonalities in successful approaches.

A useful paradigm for approaching this problem appears in Bernheim and Rangel (2009). They suggest partitioning observed decisions into those that are *suspect* or *nonsuspect*—that is, suspected of being influenced by forces that stop choice from revealing welfare-relevant preferences, or not suspected of doing so. With such a partition in hand, one can then estimate the welfare-relevant parameters (say, of a demand function or of individual utility functions) from the nonsuspect data using standard revealed-preference methods. The parameters of the model of behavioral incentive compatibility can be estimated by applying the same methods to the suspect data. Chetty, Looney, and Kroft (2009) serves as an excellent example of this approach: responses to posted prices are treated as nonsuspect, whereas responses to taxes collected at the register are treated as suspect. This partitioning generates the two different demand curves plotted in Figure 2.

When one has data on both suspect and nonsuspect choices, the framework just described serves as the default template for an empirical approach. This framework is often unavailable, however, due to the researcher determining that *all* observed decisions are suspect. In this situation, the common path forward is to seek additional data that identify the necessary features of the behavioral model. In principle, this exercise could be conducted with many forms of outside data and could even rely on estimated parameters from prior papers. However, the most common version of this approach involves designing and deploying a survey that is precisely tailored to provide the necessary missing information. This approach is well demonstrated by Kapor, Neilson, and Zimmerman (2020), who could not directly infer their probability misperceptions of interest from administrative data on New Haven school choice and thus conducted a survey that directly elicited families' beliefs about match probabilities. With such additional data in hand, estimating a model of probability misperceptions is much more straightforward.

Economists' use of tailored surveys has grown rapidly in recent years. This has both caused, and been caused by, major reductions in the logistical difficulties of deploying such a study. Researchers now have access to both user-friendly platforms for distributing surveys online and means to target the deployment of such surveys directly to the desired study participants. As a result, this tool has greater usefulness, and more and more papers are responding by using a tailored survey to fill a critical gap in field data. In such projects, the design of the survey is a key stage where creativity can be extremely rewarded: pairing the right type of data through these means can make progress possible where it would otherwise be inconceivable.

Performing Welfare Analysis

Once a researcher has estimated all the necessary model components, how should she then proceed with welfare analysis? My recommendation on this question is perhaps disappointingly uncomplicated. With the model in hand, the researcher should now directly attempt to understand the consequences of the economic decisions she set out to study. This could entail comparing welfare before or after a policy change, or comparing welfare across several different economic regimes, or comparing welfare across different values of a policy parameter to inform how it should be determined. Except for having generated the estimated models in different ways, a researcher may largely proceed as she would have if she deployed standard methods.

That said, when pursuing this welfare analysis, it is important to remember that our behaviorally informed models are still imperfect approximations. An immediate implication is that these approximations may fail if we use them to forecast behavior or welfare outside the range of situations used to estimate them. To illustrate, the results of Chetty, Looney, and Kroft (2009) suggest that sales taxes are often ignored when purchasing comparatively cheap items in a grocery store, but this finding might not hold when taxes are much higher or if the goods considered are more expensive.[4] These issues make it important to think critically about the boundaries of safe application of one's estimated model. However, it is also worth remembering that this requirement is in no way new and in no way special. The concerns above are essentially an application of the Lucas (1976) critique—that is, the concern that parameter estimates can change when the underlying policy regime changes. This critique has plagued economists regardless of their reliance on behavioral incentive compatibility.

In the course of conducting this welfare analysis, a researcher will normally wish to establish the exact role that the behavioral incentive compatibility assumption is playing. The answer to this question will of course vary across contexts, but existing research suggests a theme. Across the three studies we examined, we see clearly that individual mistakes do more than merely hurt the people who make them. This recurring finding has served as a counterpoint to the historical tendency of behavioral economists to focus their attention on the individual consequences of these mistakes. It appears that, in some cases, the consequences of behavioral influence on the broader market can be of even greater importance. In our three examples, these broader consequences included lowering the total welfare costs of taxation, preserving an insurance market that would otherwise have been significantly harmed by adverse selection, and disrupting the ability to infer utility from choices to a degree that influences which school-choice mechanism we recommend. In all three of these cases, standard welfare analysis is oriented around studying the consequences to distortions in behavior arising from optimal response to incentives. When incentive

[4] See Taubinsky and Rees-Jones (2018) for supporting evidence.

compatibility is replaced with behavioral incentive compatibility, the manner in which these distortions play out changes, thus driving the differences in the approaches.

Conclusion

The profession's tolerance of imperfectly rational "behavioral" assumptions in welfare economics has changed dramatically in recent history. Prior to the turn of the millennium, behavioral economists largely avoided engagement with full, technical welfare analysis. And indeed, such engagement would rarely have been welcomed. In the span of merely a decade or two, analysis of this variety has gone from being extremely rare to quite common, with notable examples of this analysis serving focal roles in several literatures.

As these analyses have propagated, so too has evidence on how to best pursue them. This paper has summarized commonalities in how these papers are executed and offered guidance on bringing this approach to new problems. As our focal examples illustrate, this body of work has begun to achieve the long sought-after goal of integrating behavioral economics into our most fundamental economic analyses. As the path for such research becomes more deeply trodden, I hope more and more researchers will choose to follow it.

References

Abdulkadiroğlu, Atila, Yeon-Koo Che, and Yosuke Yasuda. 2011. "Resolving Conflicting Preferences in School Choice: The "Boston Mechanism" Reconsidered." *American Economic Review*, 101(1): 399–410.

Abdulkadiroğlu, Atila, and Tayfun Sönmez. 2003. "School Choice: A Mechanism Design Approach." *American Economic Review*, 93: 729–47.

Agarwal, Nikhil and Paulo Somaini. 2018. "Demand Analysis Using Strategic Reports: An Application to a School Choice Mechanism." *Econometrica*, 86: 391–444.

Barberis, Nicholas C. 2013. "Thirty Years of Prospect Theory in Economics: A Review and Assessment." *Journal of Economic Perspectives*, 27: 1 (Winter), 173–96.

Bernheim, B. Douglas and Antonio Rangel. 2009. "Beyond Revealed Preferences: Choice-Theoretic Foundations for Behavioral Welfare Analysis." *The Quarterly Journal of Economics*, 124(1): 51–104.

Chetty, Raj. 2009. "Sufficient Statistics for Welfare Analysis: A Bridge Between Structural and Reduced-Form Methods." *Annual Review of Economics*, 1: 451–88.

Chetty, Raj, Adam Looney, and Kory Kroft. 2009. "Salience and Taxation: Theory and Evidence," *American Economic Review*, 99, 1145–77.

Dubins, Lester E. and David A. Freedman. 1981. "Machiavelli and the Gale-Shapley Algorithm." *American Mathematical Monthly*, 88: 485–94.

Gale, David and Lloyd S. Shapley. 1962. "College Admissions and the Stability of Marriage." *American Mathematical Monthly*, 69: 9–15.

Goldin, Jacob and Tatiana Homonoff. 2013. "Smoke Gets in your Eyes: Cigarette Tax Salience and

Regressivity." American Economic Journal: Economic Policy, 5(1): 302–36.
Goldin, Jacob and Daniel Reck. 2022. "Optimal Defaults with Normative Ambiguity." *Review of Economics and Statistics*, 104(1): 17–33.
Handel, Benjamin R. 2013. "Adverse Selection and Inertia in Health Insurance Markets: When Nudging Hurts." *American Economic Review*, 103 (7): 2643–82.
Harberger, Arnold C. 1964. "The Measurement of Waste." *American Economic Review*, 54(3): 58-76.
Hines, James R. 1999. "Three Sides of Harberger Triangles." *Journal of Economic Perspectives*, 13(2): 167–88.
Kapor, Adam J., Christopher A. Neilson, and Seth D. Zimmerman. 2020. "Heterogeneous Beliefs and School Choice Mechanisms." *American Economic Review*, 110, 1274–1315.
Kahneman, Daniel and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica*, 47: 263–91.
Lucas, Robert. 1976. "Econometric Policy Evaluation: A Critique." In *The Phillips Curve and Labor Markets*, Vol. 1, edited by Karl Brunner and Allan H. Meltzer, 19–46. Amsterdam: Elsevier.
Pathak, Parag A. and Tayfun Sönmez. 2008 "Leveling the Playing Field: Sincere and Sophisticated Players in the Boston Mechanism." *American Economic Review*, 98(4): 1636-1652.
Reck, Daniel. 2016. "Taxes and Mistakes: What's in a Sufficient Statistic?" SSRN Working Paper 2268617.
Reck, Daniel and Arthur Siebold. 2023. "The Welfare Economics of Reference Dependence." NBER Working Paper 31381.
Roth, Alvin E. 1982. "The Economics of Matching: Stability and Incentives." *Mathematics of Operations Research*, 7: 617–28.
Taubinsky, Dmitry and Alex Rees-Jones. 2018. "Attention Variation and Welfare: Theory and Evidence from a Tax Salience Experiment." *Review of Economic Studies*, 85(4): 2462–96.