



Foundations of Behavioral Welfare Economics

Prof. B. Douglas Bernheim
Stanford University
NBER/Sloan Behavioral Public Economics Bootcamp
May 2022

Introduction

- Economic policy analysis and policy making are increasingly drawing on insights from Behavioral Economics
- A critical component of economic policy analysis is evaluation (welfare analysis)
- Problem: Standard welfare economics *defers to choice*. Does this practice still make sense if choices can be inconsistent, biased, and so forth?
- Behavioral Welfare Economics is critical because it provides foundations for drawing normative conclusions in these settings

Outline of Lecture

- I. The behavioral critique of standard welfare economics
- II. Strategies for redesigning welfare economics
- III. Behavioral revealed preference
- IV. A general framework for choice-based behavioral welfare economics

I. The behavioral critique of standard welfare economics

I. *The behavioral critique of standard welfare economics*

A. *Premises for standard welfare economics*

- **Premise 1:** *Coherent preferences, \succeq , govern each individual's judgments about their own well-being*
 - \succeq is a well-behaved (complete, transitive) preference relation, applicable regardless of context or framing
- **Premise 2:** *Each individual is the best judge of their own well-being.*
 - Justifications: (i) arguments for self-determination in the tradition of classical liberalism; (ii) Cartesian principle that experience is inherently private and not directly observable
 - Implication: \succeq is *normative*
- **Premise 3:** *Each individual's preferences determine their choices. When they choose, they seek and achieve the greatest benefit according to their own judgment, subject to their constraints.*
 - From any choice set, the consumer selects a maximal element according to \succeq . It follows that \succeq is discoverable from choices

I. The behavioral critique of standard welfare economics

B. Classes of critiques

- ***Implementation Critiques***

- A central theme of Behavioral Economics and Psychology is that people often have difficulty making choices that advance their objectives
- Examples: people may hold biased beliefs, engage in motivated reasoning, are inattentive, misunderstand applicable principles, rely on heuristics, undermine their own objectives...
- Challenge *Premise 3*

- ***Coherence Critiques***

- Certain findings in Behavioral Economics and Psychology suggest that people do not have coherent preferences
- Challenge *Premise 1*

I. *The behavioral critique of standard welfare economics*

B. *Classes of critiques*

- A fundamental *Coherence Critique*: People don't actually have preferences that they can access – they *construct* their judgments contextually (Lichtenstein and Slovic, 2006)
 - Experiences and sensations are highly disaggregated
 - Sometimes we are called upon to render aggregate judgments, e.g., for the purpose of making a choice or reporting well-being
 - We cannot render these judgments by consulting “true preferences” or aggregate “experienced utility,” because these things don't actually exist
 - Instead, we “construct” our aggregate judgment separately within each context
 - The context of construction influences the aggregation process, for example by making various dimensions of experience either more or less salient
 - Preferences (\succsim_f) therefore depend on the “frame” in which the judgment is made

I. The behavioral critique of standard welfare economics

B. Classes of critiques

- Examples of evidence sometimes cited as support for the *constructed preferences hypothesis*:
 - Anchoring and coherent arbitrariness: Ariely, Loewenstein, and Prelec (2003)
 - Decisions with no immediate consequences are sensitive to the weather at the moment of choice: Busse et al. (2015) on automobile purchases, Meier et. al. (2016) on voting
- Cleanly differentiating context-dependent preferences from context-dependent implementation problems can be challenging

I. *The behavioral critique of standard welfare economics*

B. *Classes of critiques*

- In principle, one can also envision **Judgment Critiques**
 - Challenge *Premise 2* by asserting that people have bad objectives
 - Foundations for such critiques are not found in Behavioral Economics: in claiming that someone has bad objectives, the analyst is just expressing a difference of opinion
 - On close examination, many claims that are phrased this way are, in fact, Implementation Critiques
 - Example: one often sees the claim that people are too “present focused,” which sounds like a Judgment Critique. But the claim is that people with preferences of the form

$$U_t = u_t + \delta u_{t+1} + \delta^2 u_{t+2} + \dots$$

suffer from self-control problems that cause them to act on preferences of the form

$$U_t = u_t + \beta(\delta u_{t+1} + \delta^2 u_{t+2} + \dots)$$

In other words, they fail to implement their preferences

II. Strategies for Redesigning Welfare Economics

II. *Strategies for reconstructing welfare economics*

- Two competing paths:
 - “Fix” choice-based welfare analysis
 - Discard choice-based welfare analysis in favor of *Subjective Well-Being* (SWB) -- e.g., self-reported happiness, life satisfaction
 - These two approaches draw on different philosophical traditions, which involve disparate definitions of well-being
- SWB draws on *Welfare Hedonism*: “Well-being consists solely in the presence of pleasure and the absence of pain.”
 - Bentham, Mill
 - The general notion that welfare is exclusively a reflection of mental states is called *mental statism*.
- Choice-oriented approaches draw on *Desire Theory (Preference Theory)*: “Well-being consists in having one’s desires satisfied”
 - Means that the actual state of the world is what the individual wants it to be, and not whether the individual necessarily knows this to be true
 - A subtle variant of desire theory: the actual state of the world includes the individual’s mental state

II. Strategies for reconstructing welfare economics

A. What is welfare?

- These traditions can diverge in contexts that are central to behavioral economics (false beliefs, belief-based utility)
- The case of the oblivious altruist
 - A small town in the Arkansas experiences massive flooding, leaving many families homeless
 - Norman is altruistic, and would contribute \$100 to a relief fund if he knew about it
 - The government provides relief, paid for with taxes, including a \$100 levy on Norman
 - Norman never learns about the flood or the relief effort, and hates paying taxes
- Does the relief effort improve or reduce Norman's welfare?
 - Welfare hedonism → reduce
 - Simple desire (preference) theory → increase
 - Subtle desire (preference) theory → ambiguous

II. Strategies for reconstructing welfare economics

B. Possible paths forward

- SWB encounters other conceptual problems
 - Example (the *Aggregation Problem*): mental states are disaggregated (over time, across states of nature, varieties of states at a moment in time). Arguably, there is no aggregative mental state. We would need a principle of aggregation, and it would have to be based on something other than mental states. SWB resorts to a “linguistic” principle.
- Lecture will focus on the choice-based agenda

III. Behavioral Revealed Preference

III. Behavioral Revealed Preference

A. The basic strategy

- Many economists are reluctant to relinquish:
 - Premise 1 – The assumption that people have coherent preferences
 - Premise 2 -- The normative dictum that those preferences ought to govern welfare analysis
- However, in light of Implementation Critiques, they are willing to accept that the connection between preferences and choices is imperfect (a weakened version of Premise 3)

III. Behavioral Revealed Preference

A. The basic strategy

- Modeling strategy: supplement standard models of choice with additional elements representing the “cognitive biases” that purportedly account for those imperfections of implementation
- Elements of approach
 - Assume there is a utility function that rationalizes choices, $U(x, f)$ (sometimes called “decision utility” or “ex ante utility”)
 - Evaluate welfare according to a normative objective function, $V(x)$
 - Loosely, the difference, $b(x, f) = U(x, f) - V(x)$, reflects “bias”
- **The BRP Principle:** If enough is known about the process mapping preferences to choices (i.e., about the bias function $b(x, f)$), then one can recover both preferences and bias parameters from choice data.
- Notice: *this approach does not address the Coherence Critiques*

III. Behavioral Revealed Preference

C. Challenges facing the BRP approach

1. The challenge of specifying consumers' concerns
2. The challenge of identifying biases
3. The inflexible consistency requirement

III. Behavioral Revealed Preference

C. Challenges facing the BRP approach

1. The challenge of specifying the scope of consumers' concerns

A motivating illustration (“Norman’s lunch”):

- Suppose we ask Norman to order his lunch for a scheduled meeting one week in advance. Whether he selects a sandwich or a salad may depend on whether he is asked to decide at 1pm after he has just eaten, or at 4pm when he’s hungry. (Based on Read and van Leeuwen, 1998)
- Here, the natural assumption is that the preference domain encompasses food items, and the time at which Norman makes his choice is the frame, f , which either influences the construction of judgments or distorts the expression of those judgments into choices (e.g., because hunger disrupts cognition).
- This fact pattern admits another interpretation: Norman’s well-being depends not only on the food he eats, but also on what he orders and when he orders it. Under that assumption, a consumption bundle consists of bundles specifying both orders and meals, and there are no decision frames.

III. Behavioral Revealed Preference

C. Challenges facing the BRP approach

1. The challenge of specifying the scope of consumers' concerns

- The second interpretation of Norman's behavior suggests a variant of the BRP approach wherein the analyst expands the assumed boundaries of the consumer's concerns until all inconsistencies disappear, and then proceeds as if there are no biases
 - Example: temptation preferences, as formulated by Gul and Pesendorfer (2001), account for apparent choice reversals without assuming time inconsistency
- Two issues arise:
 1. We need an objective method for drawing lines between decision frames and elements of the consumption bundle.
 2. Depending on how we draw this line, choice-based welfare analysis may not be possible.

III. Behavioral Revealed Preference

C. Challenges facing the BRP approach

1. The challenge of specifying the scope of consumers' concerns

- **Issue #1:** Methods for drawing the line between decision frames and the consumption bundle
- Such methods must rely on non-choice evidence.
 - Evidence on choice isn't even available until one defines the boundaries of consumption bundles.
- Possible options: ask people what they care about, introspect, identify conditions that affect choice through channels other than preferences (e.g., ones that cause confusion about the options)
 - Formal methods remain underdeveloped

III. Behavioral Revealed Preference

C. Challenges facing the BRP approach

1. The challenge of specifying the scope of consumers' concerns

- **Issue #2:** Drawing the line in certain ways may preclude choice-based welfare analysis.
- *The Non-comparability Problem:* If the experience of choosing falls within the scope of consumers' concerns, then welfare is not recoverable from choice

III. Behavioral Revealed Preference

C. Challenges facing the BRP approach

1. The challenge of specifying the scope of consumers' concerns

- Conceptualization of the general non-comparability problem:
 - Choice-based welfare analysis makes prescriptions for a planner by asking what the affected consumer would choose *if offered the same alternatives*
 - In situations where consumers' concerns encompass the experience of choosing, the planner's task and the consumer's task are inherently *non-comparable*
 - Presenting the planner and the consumer with (ostensibly) the same menu does not mean the alternatives (correctly defined) are actually the same
 - If Norma's well-being depends not only on what she chooses but also on what she personally chooses to forego, her choices can't shed light on the best choice for a planner, because she personally chooses to forego nothing when the planner makes the selection for her

III. Behavioral Revealed Preference

C. Challenges facing the BRP approach

1. The challenge of specifying the scope of consumers' concerns

- A simple illustration of how seemingly sensible assumptions about consumers' concerns can lead to difficulties:
 - Norma must divide \$10 between herself and a friend
 - Norma is averse to bearing responsibility for leaving her friend with nothing when other options are available. Consequently, no matter how the task is framed, she divides the money equally.
 - However, she is inherently selfish and fervently wishes someone would take the decision out of her hands and give her the entire prize.
 - No choice problem can reveal Norma's preference. In particular, choosing between choice problems (an *avoidance problem*) won't work, because she remains responsible for the outcome.

III. Behavioral Revealed Preference

C. Challenges facing the BRP approach

1. The challenge of specifying the scope of consumers' concerns

Strategies for avoiding non-comparability problem:

1. Assume that consumers' concerns do not encompass conditions pertaining specifically to the experience of choosing (conditions of choice, as opposed to conditions of experience)
 - Maybe we can live with the assumption that conditions of choice do not matter very much, e.g., because the experience of choice is brief
2. Assume the consumer only cares about the conditions of choice under well-defined circumstances
 - E.g., we could assume that people don't care about the conditions of choice when choosing over future choice sets (Krusell, Kuruscu, and Smith, 2010, on "temptation preference")
3. Define consumption bundles in terms of mental states (Bernheim, Kim, & Taubinsky, in progress)

III. Behavioral Revealed Preference

C. Challenges facing the BRP approach

2. The challenge of identifying biases

- Implementation requires the analyst to recover V as well as U . To do so, we need to have a clear understanding of what V is.
- Potential interpretations of V :
 1. V is “experienced utility” or “ex post utility”
 2. V represents “true preferences”
 3. V is another kind of “decision utility”

III. Behavioral Revealed Preference

C. Challenges facing the BRP approach

2. The challenge of identifying biases

- **Interpretation #1** (V as “experienced” or “ex post” utility) has gained traction (e.g., Chetty, 2015), but is conceptually problematic.
- The interpretation seems to confuse welfare perspectives (appears to be slipping into welfare hedonism by equating welfare with ex post hedonic experience, which need not be the case in desire theory)
- **Problem #1**: The assumption that people derive welfare only from experience is limiting because it excludes non-experiential objectives (recall the oblivious altruist).
- **Problem #2**: There are natural and important settings in which experience cannot plausibly include the aggregate welfare evaluation V .

III. Behavioral Revealed Preference

C. Challenges facing the BRP approach

2. The challenge of identifying biases

- **Interpretation #2:** V is “true preferences,” and a “biased” choice is one that is contrary to true preferences (a common view).
- Doesn't have the same conceptual problems as Interpretation #1, since U and V are the same types of objects (ex ante desires).
- But how would we learn about “true preferences,” V , in applications?
- The dominant approach is to assume, often implicitly, that there is a decision frame for which bias is absent (i.e., f such that $b(x, f) = 0$)
- Thus, Interpretation #2 of V ultimately boils down to **Interpretation #3** (V as “decision utility” for choices that are not infected by bias)

III. Behavioral Revealed Preference

C. Challenges facing the BRP approach

2. The challenge of identifying biases

- Example: for biased beliefs, the existence of frames for which bias is absent is an *implicit* assumption

- Assume Norman chooses x to maximize:

$$U(x) = \int u(y)g(y|x)dy.$$

- The analyst infers that the objective distribution is $f(y|x)$, and evaluates welfare according to:

$$V(x) = \int u(y)f(y|x)dy.$$

- This substitution is based on an implicit (testable) assumption: in a setting where the consumer knew the objective probabilities (no bias), $V(x)$ would govern her choices.
- Notice that this is a **desire-theoretic** standard: welfare reflects the degree to which the consumer's true ex ante desires are satisfied, even if she is not aware they are satisfied

III. Behavioral Revealed Preference

C. Challenges facing the BRP approach

2. The challenge of identifying biases

-
- Unfortunately, **Interpretation #3** introduces a *Circularity Trap*: we identify bias by looking for choices that conflict with true preferences, and infer true preferences from choices that are not biased.
 - A key challenge in behavioral welfare economics is finding an escape route from this trap – i.e., a way of identifying bias without reference to preference
 - “I know it when I see it” is not a sound methodological principle
 - In practical applications, identifying the biased choices can be challenging
 - In the “Norman’s lunch” example, hunger might cloud his judgment or focus his attention
 - Framing involving the weather – does rain cause “irrational depression,” or does sun cause “irrational exuberance”?

III. Behavioral Revealed Preference

C. Challenges facing the BRP approach

2. The challenge of identifying biases

- A more consequential example: How should we evaluate time-inconsistent behavior?
- Standard model: $U_t = u_t + \beta(\delta u_{t+1} + \delta^2 u_{t+2} + \dots)$
- A widespread view: unbiased choices = “period-0” full-commitment decisions (the “long-run” criterion – ignore β)
 - Reflects the supposition that present-focus is a bias
- What principles and/or evidence support this perspective? Consider:
 - Pejorative views of present-focus are not universal (e.g., Zen Buddhism)
 - Deathbed regrets favor present-focus
 - Is the long-run criterion a reflection of “Type A paternalism”?

III. Behavioral Revealed Preference

C. Challenges facing the BRP approach

3. The rigid consistency requirement

- BRP requires the analyst to identify “biases” so that “unbiased” choices admit a coherent preference representation
- This requirement precludes the development of objective/rigorous standards and methods for identifying biases
 - No guarantee that a given set of objective principles for identifying biases will leave us with a set of internally consistent choices, and indeed no hope of success if people construct their preferences contextually
- Instead, the requirement forces one to make assumptions about bias that:
 - Lack objective support and exceed our actual understanding of choice processes
 - Recall the example of framing effects involving weather: Which type of weather causes a “bias”? If people construct their preferences contextually, the answer is potentially “none.” (Similarly for Norman’s lunch)
 - Are inherently suspect in light of the fundamental Coherence Critique

IV. A general framework for behavioral welfare economics

IV. A general framework for BWE

A. Back to foundations: revised premises

- **Premise 1:** *Each individual is the best judge of their own well-being.*
- **Premise 2:** *Coherent preferences, \succeq , govern each individual's judgments about their own well-being.*
- **Premise 3:** *Each individual's preferences determine their choices. When they choose, they seek and achieve the greatest benefit according to their own judgment, subject to their constraints.*
- How should we reformulate these premises in light of both the Implementation Critique and the Coherence Critique?
 - Based on Bernheim and Rangel (2009), Bernheim (2009, 2016, 2021), Bernheim and Taubinsky (2018)

IV. A general framework for BWE

A. Back to foundations: revised premises

- Distinguish between:
 - *Direct judgments*: pertain to outcomes we care about for their own sake
 - *Indirect judgments*: pertain to alternatives that lead to those outcomes
- Example:
 - My direct judgments may pertain to my mental states (I like some states better than others)
 - My indirect judgments may pertain to consumption goods that influence those states
- Neither behavioral critique impugns direct judgments or, by implication, correctly understood indirect judgments
 - Claims that direct judgments are flawed are just differences of opinion
 - Claims that there is some variation in direct judgments doesn't address the basis for deference

IV. A general framework for BWE

A. Back to foundations: revised premises

- **Premise A:** *With respect to matters involving either direct judgment or correctly understood indirect judgment, each individual is the best arbiter of their own well-being.*
 - **Premise B:** *When people choose, they seek the greatest benefit according to their own judgment (whether correctly or incorrectly informed), subject to their constraints.*
-

- Accommodates Implementation Critiques by allowing for incorrectly understood indirect judgments.
- Accommodates Coherence Critiques because there's no requirement that direct and correctly understood indirect judgments are mutually consistent.

IV. A general framework for BWE

B. The overall structure

- **Step 1:** Identify the scope of the consumer's concerns
- **Step 2:** Identify all decisions that merit deference (the “welfare-relevant domain” or WRD)
 - Premise B tells us that choices reflect judgments, so we retain or exclude them according to whether those judgments are correctly or incorrectly understood
- **Step 3:** Construct a welfare criterion based on the properties of choice within that domain
 - Justified by Premise A



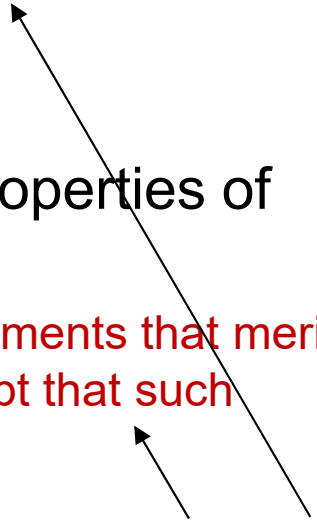
IV. A general framework for BWE

B. The overall structure

- **Relation to BRP:** One can think of BRP as involving similar steps, except we require consistency of choices after step 2, so we can conduct revealed preference analysis in step 3.
- **Key to general framework:** Devise a welfare criterion for Step 3 that accommodates inconsistencies among choices that merit deference.
 - In Steps 1 and 2, one is then no longer *required* to arrive at a WRD within which all choices are consistent
 - Free to devise objective evidence-based principles for Steps 1 and 2 that may leave us with inconsistencies
 - Creates ability to conduct welfare analysis under multiple alternative assumptions about the WRD

IV. A general framework for BWE

B. The overall structure

- **Step 1:** Identify the scope of the consumer's concerns  *Nothing new here*
 - **Step 2:** Identify all decisions that merit deference (the “welfare-relevant domain” or WRD)
 - How does one distinguish between choices that reflect correctly and incorrectly understood judgments?
 - **Step 3:** Construct a welfare criterion based on the properties of choice within that domain
 - How does one accommodate inconsistencies among the judgments that merit deference? (Under the Coherence Critique, we have to accept that such inconsistencies will exist.) 
- Our focus* 

IV. A general framework for BWE

C. Step 3: The welfare criterion

- A welfare criterion is a binary relation, W , where “ xWy ” means outcome x is better than outcome y
- A list of minimal requirements:
- **Property #1** (coherence): W is acyclic
- **Property #2** (responsiveness to choice): If, within the welfare-relevant domain, y is never chosen when x is available, then xWy
- **Property #3** (consistency with the welfare-relevant domain): If x is chosen in some decision problem with opportunity set X within the welfare-relevant domain, then x is not welfare-improvable within X according to W .

IV. A general framework for BWE

C. Step 3: The welfare criterion

- A natural candidate: the unambiguous choice relation, P^*
 - xP^*y iff the welfare relevant domain contains no decision problem in which x is available but the consumer is willing to choose y .
 - When there are choice inconsistencies, P^* will be incomplete
- **Theorem:** P^* satisfies properties 1-3. Moreover, it is the unique binary relation satisfying these three properties.
 - Implication: we need to build welfare analytics around P^*
- In cases where the criterion isn't sufficiently discerning, deeper Step 1 & 2 analysis may be fruitful

IV. A general framework for BWE

C. Step 3: The welfare criterion

- Application is intuitively straightforward
 - Example: depending on framing, Norman always chooses a mug over \$4, and always chooses \$5 over a mug, but decision is frame-dependent in between \$4-\$5
 - In that case, we can say that the equivalent variation associated with having the mug is between \$4 and \$5.
- Analytic tools of standard welfare economics extend naturally: equivalent & compensating variation, consumer surplus, etc.
- Tools for aggregation also extend (e.g., generalized Pareto optimum, aggregate equivalent variation & compensating variation)

IV. A general framework for BWE

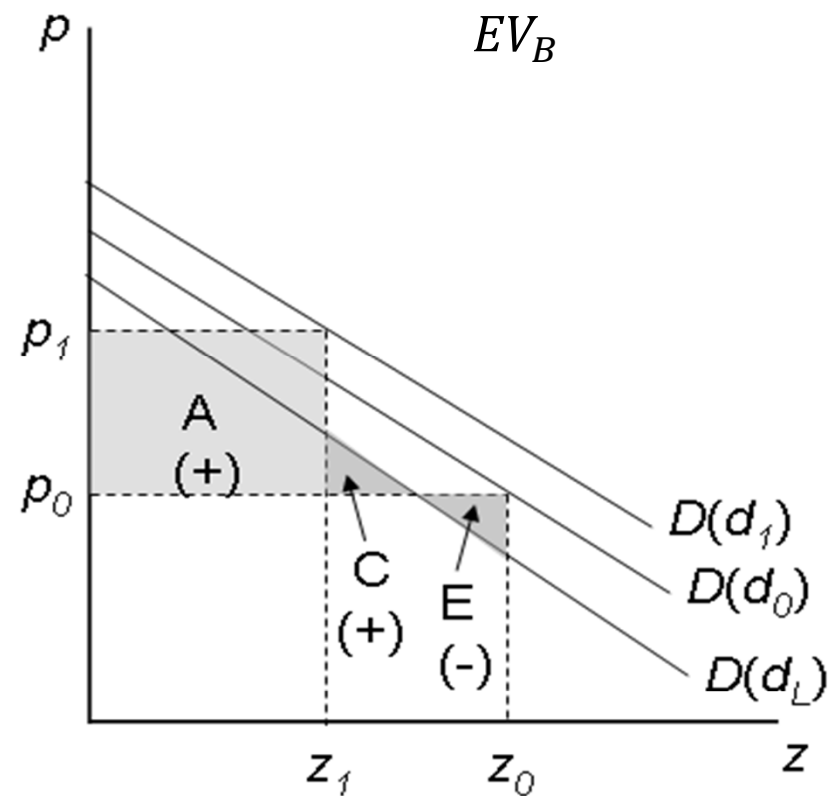
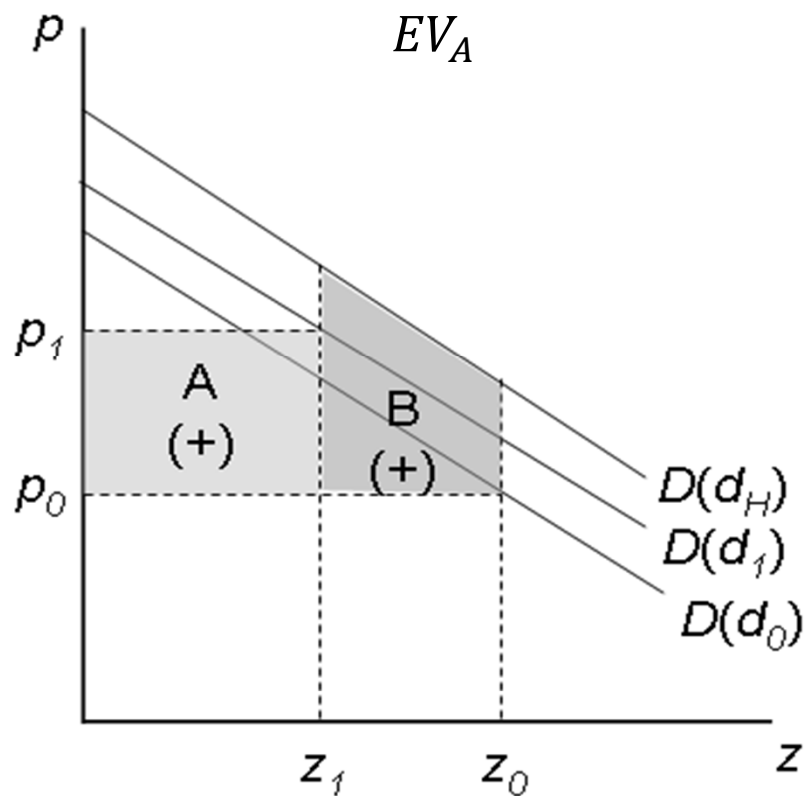
C. Step 3: The welfare criterion

- Two notions of equivalent variation for replacing choice problem G_0 with choice problem G_1 (similar for CV)
 - EV_A is the *smallest* increment to income in choice problem G_0 such that the bundle obtained with G_0 is unambiguously chosen (P^*) over the one obtained with G_1 .
 - EV_B is the *largest* increment to income with G_0 such that the bundle obtained with G_1 is unambiguously chosen (P^*) over the one obtained with G_0 .
- $EV_A \geq EV_B$, and they coincide when the choice mapping is consistent (WARP)
- One can say that the policy change is unambiguously worth at least EV_B , and no more than EV_A
- With no income effects, EV and CV concepts can be captured by standard demand curves, analogously to Marshallian consumer surplus

IV. A general framework for BWE

C. Step 3: The welfare criterion

- Generalized consumer surplus for a shift from (p_0, d_0) to (p_1, d_1) :



IV. A general framework for BWE

D. Step 2: Identifying the welfare-relevant domain

- Our objective in Step 2 is to identify and remove choices that reflect incorrectly informed indirect judgments
 - An incorrectly informed indirect judgment is one in which the decision maker mischaracterizes either the available actions or their consequences (*Characterization Failure*)
- We escape the Circularity Trap by referencing aspects of *decision processes* rather than *preferences*.
 - The fact that Step 2 precedes Step 3 precludes circularity (one cannot rely on the results of Step 3 in Step 2)

IV. A general framework for BWE

D. Step 2: Identifying the welfare-relevant domain

General strategies for identifying Characterization Failure:

1. Investigate whether, for a given class of decision problems, people make correct inferences about the available actions and their consequences
2. Investigate whether, for a given class of decision problems, people are equipped to make correct inferences about the available actions and their consequences
 - Are they aware of and knowledgeable about applicable principles?
 - Do they deploy applicable principles and relevant information?
 - Do they deploy cognitive processes involving attention, memory, forecasting, etc., as required to infer options and consequences?

IV. A general framework for BWE

D. Step 2: Identifying the welfare-relevant domain

- What if it turns out that Step 2 yields a WRD that is “just right” in the sense that it is both comprehensive and internally consistent, rather than “too large” (i.e., still contains inconsistencies)?
 - P^* specializes to standard revealed preference
 - We thereby arrive at the third interpretation of V given previously: it is just a function that rationalizes choices (“decision utility”) within a special subset of decision frames
 - The framework therefore provides a true generalization of the BRP approach