Problems with Hypothetical Pareto Improvements

Toby Ord*

Introduction

The most well established method in economics for comparing two options that affect multiple people is Pareto-superiority:

> *Pareto-superiority*
>
> An option is Pareto-superior to a second option if it is better for at least one person and worse for no-one.

The main virtue of Pareto-superiority is that it is a way of comparing options that affect multiple people without needing to compare the amount of gains for one person to the amount of losses for another. It thus sidesteps many tricky ethical questions about interpersonal value comparison. The price it pays for this is that it provides only a very weak partial ordering: in almost all non-trivial cases of policy choice one option is better for some and worse for others. Thus in almost all nontrivial policy choices Pareto-superiority offers no guidance.

Furthermore, while it is plausible that Pareto-superiority of one option over another is a sufficient condition for that first option being better all-things-considered, it is implausible as a necessary condition. For instance, a world where everyone has excellent flourishing lives is obviously better all-things-considered than a world where one of these people has a life that is slightly more flourishing than this, while everyone else lives in eternal torment, but since it is worse for someone, it is not Pareto-superior.

Hypothetical Pareto-superiority was invented to allow more policies to be compared while still avoiding interpersonal comparisons of value. Two separate criteria were created to try to achieve this:

> *Kaldor-superiority*
>
> An option is Kaldor-superior to a second option if those who would be better off under it could compensate those who would be worse off under it such that at least one person ends up better off and no-one ends up worse off.

> *Hicks-superiority*
>
> An option is Hicks-superior to a second option if those who would be worse off under it could not compensate those who would be better off under it such that at least one person ends up better off and no-one ends up worse off.

* University of Oxford.

There are various problems using each of these criteria on their own (for example you can find cases known as Scitovsky reversals where option X is superior to option Y and vice-versa). Some of these problems can be avoided if one moves to a combined criterion, so this is what is typically done.

> *Kaldor-Hicks-superiority (aka Hypothetical Pareto-superiority)*
>
> An option is Kaldor-Hicks-superior to a second option if it is both Kaldor-superior and Hicks-superior.

Kaldor-Hicks-superiority is often thought of as moving a system closer to Pareto-superiority. It is the cornerstone of Benefit-Cost Analysis (BCA), which is a very widely used measure of which policy to choose within government departments. BCA uses it as a cardinal measure for the benefits of a policy, looking at exactly how much more the winners would be prepared to pay to have the option and subtracting how much the losers would need to be compensated. This quantity is often called the 'net benefit' of a policy (when compared to a status quo). I use scare quotes to remind the reader that a 'net benefit' does not imply an overall increase in value or even an increase in resources — it is just a very leading technical term.

When economists are being careful, they are aware that there are a number of problems with using this measure to guide policy choice. There is no problem that I will present here which is unknown to all economists. However, many economists are unaware of some of these issues, a great many people who apply these principles in making decisions in government departments are unaware of some of these issues, and even when people are aware of them, I do not think that they take them seriously enough.

Technical problems

While the move to Kaldor-Hicks-superiority avoids the problems of two options being superior to each other, it still does not guarantee transitivity. It is possible to have three options (A, B, C) such that A is KH-superior to B, which is KH-superior to C, but where A is not KH-superior to C. Transitivity is a required condition for a relation to be an ordering, so KH-superiority is not an ordering (not even a partial ordering like Pareto-superiority). Moreover, this also shows that 'net benefit' is not a quantity. KH-superiority is more like *to the east of* than it is like *longer than*, while 'net benefit' is more like *easterlyness* than like *length*.

This does not show that KH-superiority is completely incoherent. It could still be a sufficient condition for one outcome being better than another, but it cannot be a necessary condition. It does show that one needs to take great care when using it though, as it is not clear that (for example) you can perform numerical operations on 'net benefits'.

Distributional problems

The most well-known problem with KH-superiority is that it is blind to the distribution of benefits. For instance it does not care how income is distributed in a society, as the incomes could always be equalized in the hypothetical transfer. This is quite widely recognized as a limitation and economists often admit that KH-superiority is an implausible criterion for better-all-things-considered. However, they typically say that we can separate questions of 'growing the pie' (economic efficiency) from 'dividing the pie' (distribution). BCA and 'net benefit' is about the former and is not trying to address the latter. They admit that a full account of policy making (or betterness all-things-considered) will have some extra steps to deal with distribution but that this is not the role of BCA, and is possibly not the role of economics at all. Instead it is seen as a role for democracy or ethics.

There are several reasons given as to why things should be cut up like this. One is that while each policy preferred by BCA may have some winners and losers, the fact that it creates so much wealth means that over many choices who wins and who loses will balance out, but the wealth of everyone will increase. Sadly it is not so, since BCA has a bias towards satisfying the preferences of the rich for they are willing to pay more for any benefit and willing to pay more to avoid harms. Demonstrating a 'balancing out' would require very unrealistic assumptions, such as that everyone has equal incomes.

Another reason given is that economics is a science and so should be restricted to the descriptive rather than the normative. On this view, the BCA is merely making a descriptive claim about whether the winners could compensate the losers and this is all it can do. Anything else is left for non-economists. A big problem with this is that it is disingenuous. The terms 'net benefit' and 'benefit-cost analysis' both use the normative term 'benefit' and derive considerable influence from this. Even as a descriptive measure, it is just one of many. Why not look at what would make people happier? Or what maximises a function of wins and losses that is weighted by the incomes of the winners and losers? Choosing to focus on just one descriptive measure in policy making is itself a normative choice.

Furthermore, why should we separate out the choice into these two parts and maximise the first part? Even if there were a proof that all socially optimal policies could be reached by such a method, it wouldn't follow that we should perform step one of this plan if we know that step two won't be performed. For example, it might maximise 'net benefits' to charge the same price to everyone for a service, but if we know that distributional needs are not going to be fully factored in through taxation or other steps, it could be better all-things-considered to charge a lower price to poorer people.

Sometimes it shrinks the pie instead of growing it

It is easy to see that increasing 'net benefit' can sometimes 'shrink the pie'. This is because it weights benefits for rich people more highly than for poor people since the rich people have a higher willingness to pay for them. Let's take an example where a gain is intuitively much smaller than a loss. Suppose someone punches someone else in the face for a mild amount of pleasure. Intuitively the pleasure is outweighed by the pain. But if the person who would like to punch is very rich, and the person to be punched is very poor, then the rich person is willing to pay more than enough to be able to compensate the poor person and the option where they punch the poor person is KH-superior. Suppose the rich person is a billionaire who is prepared to pay $1,000 for the opportunity and the poor person is a beggar who would prefer $100 to avoiding the punch. In this case, the option where the beggar is punched is considered to have $900 of 'net benefit' over the status quo.

If compensation was paid (say $500), then this would be a trade and would be Pareto-superior to the status quo. While icky, it is not clear that it is worse than the status quo, since the poor person did value the money more than avoiding the pain (my argument does not depend on whether or not this trade would be worse than the status quo). But if the rich person punches the poor person in the face with no compensation, the outcome is quite clearly worse than the status quo, and was a shrinking of the pie of social resources rather than a growing of it.

What goes wrong for KH here? The problem is that the Pareto-superior trade has two components: a punch, which lowers overall welfare, and a transfer of money from the very rich to the very poor, which greatly increases overall welfare. Together, they make a Pareto improvement, and arguably an improvement overall. However, BCA just tests for this overall improvement, but doesn't test whether the first component was increasing welfare or whether it was reducing it. In this case just making the transfer would be the best option, and just punching would be the worst option.

The example with the punch was unrealistic, but chosen to memorably illustrate the point. There are many real-world examples. For instance in international climate policy, using BCA can lead to very bad conclusions due to the disparity of wealth between the parties. There are many cases where factory owners in rich countries could lose a lot of money if they have to reduce emissions, while people in poor countries would be prepared to accept the emissions-induced risk of dying in order to be compensated with that much money. If there were a trade and the factory owners compensated for their pollution, that might be a good outcome, but BCA will recommend allowing the rich factory owners to pollute without compensating just because they *could* compensate the poor people. This will be allowed even when the gains to the factory owners are quite clearly smaller than the risk of death. Luckily some economists in the climate policy arena reject BCA for these reasons and use a version similar to the one I propose below.

Effectively makes interpersonal value comparisons (and bad ones)

The main virtue of Pareto-superiority over the methods that went before it was that it avoided making interpersonal value comparisons. However, KH-superiority effectively makes interpersonal value comparisons, and it makes ones that are much less plausible. Consider the option of taking $1 from person A and giving it to person B, while also creating a penny of value for person B. This is KH-superior to the status quo as the winner could compensate the loser and keep the penny. If a penny of value was destroyed instead of added, it would be KH-inferior. This effectively means that you can perform interpersonal value comparisons.

It is an interpersonal value comparison to say that the (social) value of an extra $1 for person A is the same as the (social) value of it for person B, and this is the same as claiming that it is equally good to give the dollar to either person. KH-superiority is effectively making this latter claim. However, it is implausible that a dollar is worth the same to everyone regardless of their existing wealth. Indeed it is a commonplace that each dollar is worth less to you when you are richer compared to when you are poorer. However, KH-superiority effectively insists that all dollar-transfers are neutral. Moreover it insists that transferring $1,000 between any two people is neutral, or transferring $1 from a million people to one recipient is socially neutral. This is very implausible!

There are better alternatives

If we are going to effectively allow interpersonal comparisons, why not do it in a much better way? If we know the incomes of the individuals, and we know the curve relating income to personal utility (which we have experimentally measured from people's preferences over gambles), then we can do much better. We can make the assumption that people get the same utility from the same income, then do our calculation of gains and losses in utility space instead of in dollars. For example, if utility is the square root of the dollars of income, and a policy would make people on $25,000 PA lose $5,000 PA, and an equal number of people on $10,000 PA gain $4,000 PA, we can see that the loss of utility is $\sqrt{25{,}000} - \sqrt{20{,}000} = 158 - 141 = 17$, and the gain is $\sqrt{14{,}000} - \sqrt{10{,}000} = 118 - 100 = 18$, so the transfer would be considered an improvement.

This method requires some extra information to be useful, but it does not really require any more assumptions, as the BCA method effectively makes just as many interpersonal utility comparisons, but does so with the (very implausible) assumption of a linear relationship between income and utility for everyone.

This method is basically the method that was used by the early economists before the move to Pareto improvements and appears to be much better than KH-superiority overall. As mentioned earlier, the main criticism of it was its reliance on interpersonal value comparisons, but KH-superiority effectively has these, as does any distributive principle that economists would advise us to use in conjunction with KH to smooth over some of its biggest problems.

Why it matters

BCA is *very* widely used and allocates billions of dollars annually, affecting the lives of billions of people. It is thus critically important to do it as well as possible. In theory, applying BCA with some additional distributive principles, could perform as well as my proposed replacement — particularly as the distributive principles and how to combine them with BCA have not been specified, so they could theoretically just transform BCA into the replacement method. However, in practice the distributive part relies on nebulous principles and is given nebulous weighting. It thus frequently has little or no effect, and when it does have an effect, it is not clear that it has the right effect. Moreover, the base tool of BCA tries to maximise something that is not even a quantity and so cannot be maximised.

Since BCA makes such a large impact on the world, it is critical that we do it as well as possible, and we could do much better than the current method based on hypothetical Pareto-improvements.