COMMENTARY

The Indomitable in Pursuit of the Inexplicable: The World Development Reports' Failure to Comprehend Economic Growth Despite Determined Attempts, 1978–2008

William Easterly

The intellectual tragedy of 30 years of *World Development Reports* (*WDRs*) is that they never accepted the reality of the great unpredictability and uncertainty of economic growth in the short to medium run. The *WDRs* keep trying to find ways to raise growth in the short to medium run when the economics profession does not have this knowledge. They seek to explain short-term fluctuations in growth when there is no evidence base for such explanations. As a result, they fall prey to many of the classic heuristic biases about randomness (à la Kahneman and Tversky), including frequent use of circular reasoning, and they lose the opportunity to carry on a fruitful debate about the best way to handle this uncertainty and to make development more likely in the long run (Gilovich, Griffin, and Kahneman 2002; Kahneman, Slovic, and Tversky 1982).

What is the state of our knowledge about growth? First of all, country growth rates are not persistent over time, which was documented as long ago as Easterly, Kremer, Pritchett, and Summers (1993). High growth is

mostly transitory, reverting to the global mean in the following period. This finding was bad news when most of the candidate explanations of growth were very persistent country characteristics. Of course, there could have been time-varying variables that explained the time-varying element of growth. Unfortunately, the second characteristic of our growth knowledge is that we have failed to identify any such robust time-varying variables (or for that matter any robust persistent variables). Levine and Renelt (1992) established this failure convincingly early in the growth literature. It further showed itself in the 145 different variables found to be "significant" in growth regressions with fewer than 100 observations (Durlauf, Johnson, and Temple 2005). The last hope was Bayesian model averaging to identify the small number of variables that were robust in most regressions (Doppelhofer, Miller, and Sala-i-Martin 2004). Even this hope vanished recently when Ciccone and Jarociński (2008) showed that Bayesian model averaging gave completely different "robust" variables for different equally plausible samples (World Bank versus Penn World Tables or successive revisions of the Penn World Tables).

In defense of the *WDRs*, the economics profession was also slow to admit the inexplicability of growth fluctuations. However, a wide spectrum of economists has by now conceded we don't know how to raise growth in the short to medium run (Easterly 2001; Lindauer and Pritchett 2002; Harberger 2003; "Barcelona Development Agenda" 2004;[1] Rodrik 2006; Solow 2007; Spence Commission 2008).

A random effects regression on the panel of per capita growth rates from 1960 to 2005 reveals that only 8 percent of the cross-time, cross-country variation in growth is due to permanent country effects; the other 92 percent is transitory (which is equivalent to stating the lack of persistence of growth rates identified in Easterly, Kremer, Pritchett, and Summers 1993). The transitory does not have to be mechanically "random" in the sense of coin-flipping; it could well be one-off movements caused by human action. It could be an entrepreneur finding a "big hit" in exports, like cut flowers in Kenya or garments in Bangladesh; it could be a smart policy move that

1. The "Barcelona Development Agenda" is a consensus document resulting from a meeting of economists in Barcelona, Spain, in 2004. Signatories of the document include Olivier Blanchard, Guillermo Calvo, Stanley Fischer, Jeffrey Frankel, Paul Krugman, Dani Rodrik, Jeffrey Sachs, and Joseph Stiglitz.

was in the right place at the right time; or it could be a bubble caused by an information cascade or other kinds of herding. On the negative side, it could be a dramatic mistake by a policy maker or a private entrepreneur. Still the transitory might as well be random in the sense that we cannot usually explain or replicate what just happened.

Hence, many of the classic Kahneman-Tversky heuristic biases about randomness have played themselves out in *WDRs*.[2] Take, for example, the fallacy of the "hot hand," when a basketball player makes a string of baskets in a row. The hot hand bias is to falsely conclude that the player's skill has temporarily moved to a higher level, whereas actual calculation shows that a player is no more likely to make the next basket after a hot streak than at any other time. The problem is that we expect randomness to show up as alternating hits and misses when in fact it often displays streaks of hits. Another way of stating this fallacy is Kahneman and Tversky's sarcastically named "law of small numbers" (Kahneman, Slovic, and Tversky 1982). In the case of the *WDRs*, we falsely draw conclusions about how to achieve superior long-run performance from too small a number of observations, without allowing for the large role of transitory factors in a small sample. The small numbers refer both to a small number of "successes" and a small number of annual observations (even 25 years may not be long enough, as will be discussed).

WDRs abound with statements reflecting this fallacious viewpoint, as summarized by Yusuf:

> If [China and India] can rack up rates of investment and growth that are the envy of the world under the most makeshift of institutional conditions, need other countries more attuned to the market strive after greater perfection? China was growing when it had few if any market institutions; as its institutional structure has strengthened, it has continued growing with investment serving as the principal driver without a clear relationship running from the specifics of institution building to growth.

China and India definitely reflect some genuine success, but their sudden shift upward in growth is also bound to reflect some inexplicable, transitory factors that do not help us understand success (and it is even worse to break up their performance into subperiods, as with China in the last sentence).

2. A wonderfully entertaining summary of this and other related research is a recent book for nontechnical audiences by Mlodinow (2008).

One systematic way of showing the hot hand fallacy at work is by simulating a mechanical procedure to identify "success." The example I use is not from *WDRs* but from the Spence Commission (2008); however, the *WDRs* (as shown by the quotes above) definitely do informally what the Spence Commission did more formally, so this example is just a way to formalize a comment on the *WDRs*' worldview.

The Spence Commission identified "success" as (essentially) any 25-year period of gross domestic product (GDP) per capita growth above 5 percent.[3] This procedure sounds like a pretty good bet, but in fact it was very likely to pick up a large element of transitory performance for two reasons:

1. Selecting on high values of the growth outcome will very likely include large positive realizations of the transitory component. This problem is all the more likely because the permanent component of growth outcomes exceeds 5 percent in only 1.8 percent of realizations (whereas the temporary component will exceed 5 percent by itself in 26 percent of realizations).
2. Selecting on the time period (*any* 25-year period out of a 45-year sample from 1960 to 2005) further biases the episodes toward those that had large positive transitory outcomes. The time period is selectively biased to be one that started and ended so as to include a large number of large positive transitory outcomes.

A Monte Carlo simulation based on the parameters from the random effects regression shows that the Spence Commission's definition of "success" will occur in about 9 percent of countries, which is far more than the 1.8 percent of countries that have a genuine permanent country growth above 5 percent (granted the assumptions about the permanent and transitory components being normally distributed). In the event, the Spence Commission found 13 "success stories."[4] Interestingly, India did not make

3. I say "essentially" because the commission inexplicably used total GDP growth rather than per capita growth. Its criterion was GDP growth above 7 percent, so with population growth usually about 2 percent, I convert this criterion to a per capita growth criterion of above 5 percent.

4. I did 25,000 runs of per capita growth in countries for 45 years, in which growth is the sum of two orthogonal components: a normally distributed permanent component $N(0.0176438, 0.0155495)$ and a normally distributed transitory component $N(0, 0.0506495)$. The means and standard deviations are taken from the random effects regression over 1960 to 2005 of all countries

it on the Spence exercise, suggesting that informal discussions of success stories are even looser than the excessively loose Spence criterion.

The Spence Commission spent a lot of time analyzing these high-growth countries as if they completely reflected fundamentals. However, the other bad news about the bias toward including a large transitory element is that this procedure will likely not even pick the right countries. The same Monte Carlo simulation reveals that about 37 percent of the countries that are in the top 9 percent according to the Spence criteria are *not* in the top 9 percent of permanent country growth rates. The Spence Commission successes (just like the *WDR* success story analyses)—even as they are carefully being picked apart to discern their innermost secrets—are bound to include some ringers that just got lucky.

Why is such flawed analysis pursued by such talented and well-trained economists? Yusuf notes with frustration that "even with good policies, the growth of the typical developing country rarely climbs much above 3 to 5 percent per year [1 to 3 percent per capita]." Yusuf notes that this figure "is impressive by historical standards, but countries in a hurry to catch up aspire to faster rates of growth." The Spence Commission and the *WDRs* just cannot accept that 5 percent per capita growth is rare (expected to occur in 1.8 percent of the sample). It is easy to see the appeal of a definition that makes this yearned-for outcome 4.8 times more likely, and so economists are often willing to overlook that this increased likelihood is likely spurious.

So we see "growth booms" as attainable because we think they reflect an intentional shift in the country's fundamentals upward, which could be replicated elsewhere. Again, this assumption could possibly be right, and we could have confirmed it if we had achieved any success in explaining cross-time variations with some variables capturing fundamentals—but we have not done so. Or the *WDRs* could successfully be doing qualitative analysis that would help identify ways to trigger a growth boom. However,

with complete data so as to have a balanced panel (95 countries). The Spence Commission found 13 "success stories," but the commission does not say how large its sample of countries with the necessary data was. Thirteen would be 9 percent if the sample was 144 countries, which sounds a little too high for countries having complete data. Of course, one run of 100 or so countries is not large enough to give a precise estimate of the percent likelihood of "success"; such a small sample estimate could vary considerably around the expected value computed from a large value of Monte Carlo simulations.

Yusuf's review shows instead the frequent changes in messages, the sloppy vagueness of explanatory factors, and a complete lack of success stories in replicating growth booms through expert advice in the *WDRs*. It seems like the hot hand fallacy may instead explain our unproductive fascination with growth booms.

This heuristic bias is so hardwired into us as humans that we actually do worse than rats on the hot hand fallacy. In a classic laboratory experiment, subjects were shown a light that flashed either red or green. They were allowed to watch for a while and then were asked in successive rounds to predict the next flash. The experiment was rigged so that red was randomly flashed twice as often as green, although the subjects were not told so. The rats pursued the optimal strategy of always guessing red. The humans did not. The humans thought they perceived occasional "hot streaks" of green and would then guess green. As Mlodinow (2008) says "humans usually try to guess the pattern, and in the process we allow ourselves to be outperformed by a rat."

Another heuristic bias is called the "halo effect." This effect is the well-documented tendency (verified in many psychology experiments in the laboratory) to assume that an individual who excels on one dimension will also have superior talents on other dimensions (as subjectively evaluated by the observers in the experiments, for which there is no factual basis whatsoever by the design of the experiment).[5] So, for example, we expect our successful male politicians to also be good husbands (despite abundant evidence to the contrary). And *Fortune* magazine's annual ranking of the World's Most Admired Companies ranks companies on eight very different dimensions, which are all suspiciously correlated with the company's latest financial performance and with each other. So Cisco Systems was highly rated on quality of management, quality of people, innovativeness, and so forth in 2000, when its stock value was high. When the stock collapsed after 2001, observers suddenly detected that every dimension got worse at the same time: the same management and people had overnight become low quality and not innovative (Rosenzweig 2007: 61–62).

One particularly remarkable laboratory finding came from an experiment in which subjects observed two people executing a task. The

5. This effect is also the subject of an excellent book for nontechnical audiences (Rosenzweig 2007).

experiment had been carefully rigged so that the two people's performance was equal. The subjects were told that one of the two people would receive a large payment and that this assignment would be *random*. The subjects were then asked to describe the performance of the two agents. Despite the subjects' knowledge that the payment was random, they gave superior marks on multiple performance attributes to the agent who received the payment.

In the *WDRs*, a country that excels in achieving high growth is assumed to also excel in having wise leaders, good institutions, entrepreneurial citizens, and so on. The latter characteristics are hard to measure objectively, so these subjective assumptions are hard to prove or disprove. Then, to go from the halo effect to pure circular reasoning, we conclude that these wise leaders, good institutions, and entrepreneurial citizens explain the high growth.

Perhaps the worst single offender with respect to the halo effect and circular reasoning in the *WDRs* was the introduction of the concept of the "investment climate." This concept absorbed one entire *WDR* and yet lacked any theoretical definition or any agreed-upon measurement. Something so vague is bound to be seen wherever good outcomes are happening and then flexibly deployed to "explain" success. Yusuf diplomatically acknowledges these problems: "Nick Stern, the Bank's chief economist from 2000 to 2003, was instrumental in making the assessment of the investment climate in member countries an integral part of the Bank's economic analysis of countries. His conception of the determinants of this climate was sweeping ...." It was so sweeping as to use what Yusuf politely calls an "eclectic selection of evidentiary material." Yet the appeal of circular reasoning through the halo effect still holds: "Did Botswana, Chile, China, India, and Mauritius as well as the East Asian economies achieve growth mainly by mending the investment climate ...?"

The halo effect contaminates the endless and increasingly useless analysis of the East Asian success stories. Hong Kong, China; Taiwan, China; the Republic of Korea; and Singapore are very unlikely to be ringers; they almost certainly represent genuine long-run success on growth rates. Yet the halo effect falsely anoints every single aspect of these countries as also being ultra-exceptional and then jumps to the unwarranted conclusion that every such factor contributed to the remarkable success. The successful

East Asian characteristics are subjectively chosen, and it is even worse that they seem to keep changing with whatever is the latest fad in development thinking. Yusuf states:

> East Asian economies, by virtue of their successful growth performance, became the ones to emulate. The message distilled from their experience was that market-guided industrialization within the milieu of a relatively open economy could result in rapid growth if industries were able to compete in export markets.
>
> . . .
>
> [T]he success of a China or a Korea or a Singapore rested on the state's readiness to trim the public sector, encourage private enterprise, and build market institutions, but in each case, the state has remained large, powerful, and interventionist. Directly and indirectly, the public sector encompasses a major share of GDP.
>
> . . .
>
> Everyone can see that market institutions in successful East Asian industrializing countries are at best functional and at worst weak and minimally supportive. The interesting issue is how an assortment of institutions of varying capabilities and degrees of maturity can, with the help of a strong developmental state, produce good results using the local knowledge that policy makers surely have.

Then, to make things yet worse, we jump to conclusions from an even smaller number of recent observations in which the Gang of Four slowed down:

> Other high-performing countries in East Asia have seen their growth performance flag while their institutions have matured, albeit slowly. However, all these economies have also witnessed a decline in investment and a partial withdrawal of the state from the forefront of economic decision making.

As if this were *still* not bad enough, the analysis of the few top performers is contaminated even further by yet another selection bias: the survivor bias. Suppose that a set of drivers was going from New York to Washington, D.C., driving Lamborghinis at 150 miles per hour down I-95. We are in Washington and interview the Lamborghini drivers who arrive. We wax ecstatic at the drivers' trip to Washington in under two hours (compared with the usual minimum of four hours), their willingness to take bold risks, and the overall superiority of the speeding Lamborghini drivers to the other plodding drivers on I-95. Because we observe only the ones who arrive in Washington, we are unaware that many (plausibly a large majority) of the Lamborghini speedsters were pulled over and arrested for

reckless driving and never made it to Washington, not to mention a few who were killed or maimed in traffic accidents because of their insanely risky driving. So on average, the hockey moms driving minivans, who arrive in Washington in five hours or so, outperformed the Lamborghini drivers. Our conclusion that going 150 miles per hour in a Lamborghini is a formula for success in getting to Washington is false; we were led astray by survivor bias.

We induce a survivor bias when we analyze only the top "success stories." I doubt very much that the success of the Gang of Four is entirely explained by survivor bias. But this example does show the risks of praising every aspect of the experience of the Gang of Four. Some strategies may have been very risky, and by concentrating only on the success stories, we miss the experience of other countries that may have followed the same strategy and crashed and burned. Survivor bias makes the whole methodology of obsessively dissecting every aspect of the success stories very suspect. The remedy is simple: to assess the growth payoff from factor *X*, we should study *all* countries—both those that had factor *X* and those that did not—and ask, "What was the average payoff?" So take, for example, the conclusion sometimes reached that the Gang of Four's success is due to authoritarian leaders pursuing industrial policies. But the track record worldwide of dictators picking winners is very poor, so why are we so sure that this factor contributed to the success of the Gang of Four? And even if it did, which is basically nonfalsifiable, why do we think it is replicable elsewhere—that finding which is most relevant and *is* falsifiable?

Of course, the general enterprise of assessing all possible factor *X*s to find the secrets to growth success has not been helpful either (see the previous discussion of growth literature), but at least this exercise was not contaminated by survivor bias. We have still learned something from the failure of growth regressions: that there is no universal factor *X* that works everywhere to reliably raise growth—because if there had been, it surely would have shown up as a robust determinant of growth in our extensive effort at cross-country regressions.

On a more positive note, how *should* we deal with a world where there is so much uncertainty about growth determinants? Despite this uncertainty, a substantial number of countries (Australia, Japan, the Gang of Four, and countries in Europe and North America) have already achieved a high level of per capita income, which must reflect good average growth

performance over some suitably long period. The problems with randomness get progressively alleviated the longer we make the period of analysis. Studying the *level* of per capita income rather than growth rates as a measure of success or failure is one way to focus on the long run. The *WDRs* have been forced by the peculiar conventions of development economics to exclude most of the countries that actually succeeded the most at development, and so they rarely invoke any lessons from the long histories of countries that are now rich (except for the Gang of Four), as compared with those that are still poor. In contrast, a slew of papers that were published in top journals in economics studied levels for the whole sample and attributed development success to long-run factors such as property rights, democracy, trade openness, and technological creativity. These papers have their own problems resolving correlation and causation, but they are still clearly superior to the methodology of the *WDRs*; the latter have been led fatally astray by glaring biases in the treatment of transitory components of volatile short- to medium-run growth rates.

Perhaps one way to unify the findings of the levels regressions—a theoretically appealing way to understand how systems can handle vast short-run uncertainty—is to hypothesize that systems that respect individual rights do the best in the long run on economic development. Such individual rights include property rights, rights to dissent from prevailing conventional wisdom, rights to trade whatever with whomever you want, rights to enter new industries and start up new firms, rights to advocate new political directions, and so on. The theoretical appeal of this hypothesis is that individual rights can handle systemic uncertainty by exploiting individuals' superior localized knowledge and powerful incentives to solve their own local problems, which will lead to superior performance even if no policy maker at the top knows how to raise growth rates.

This possibility is obviously just the beginning of such a discussion, and this brief discussion is a long way from confirming this or any other hypothesis. The sad thing about the *WDRs* is that they missed out on such fruitful and deeper long-run discussions about the best systems for achieving development under uncertainty by diverting all their energies to a futile attempt to find patterns in this uncertainty. Are our heuristic biases, like those described here, so strong that future *WDRs* will continue this tragic intellectual failure? As usual, it is hard to predict.