EJW
ECON JOURNAL WATCH
Scholarly Comments on Academic Economics

Econ Journal Watch

Scholarly Comments on Academic Economics

Volume 9, Issue 3, September 2012

COMMENTS

ECONOMICS IN PRACTICE

WATCHPAD

EJW
ECON JOURNAL WATCH
Scholarly Comments on Academic Economics

ECON JOURNAL WATCH 9(3)
September 2012: 170-185

Big Questions and *Poor Economics*: Banerjee and Duflo on Schooling in Developing Countries

James Tooley[1]

LINK TO ABSTRACT

Can we avoid development's "big questions"?

In their widely acclaimed 2011 book *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*, MIT professors Abhijit Banerjee and Esther Duflo argue that too many experts, like Jeffrey Sachs on the left and William Easterly on the right, get "fixated on the 'big questions'" such as "How much faith should we place in free markets?" and "Does foreign aid have a role to play?" (Banerjee and Duflo 2011, 3). Instead, we should move away from this blinkered left-right debate about development and focus on real problems like "how best to fight diarrhea or dengue" (3). With evidence now available, including datasets from 18 poor countries and data from randomized controlled trials (RCTs), "it is possible to make very significant progress" fighting global poverty "through the accumulation of a set of small steps, each well thought out, carefully tested, and judiciously implemented" (15).

Poor Economics covers most of the topics you'd expect in a development expert's portfolio, including food, health, family planning, microfinance and microcredit. Each of the areas is contextualised with stories of the realities of lives of the poor, with evidence adduced to support or dismiss particular policy proposals. All of the chapters are interesting and challenging, but in this article I

1. Newcastle University, Newcastle upon Tyne, UK NE1 7RU.

treat only their discussion of education, in chapter 4, "Top of the Class" (72-101), framing this within the context of their general approach found in chapter 1, "Think Again, Again" (1-16).

I favour much of their general approach. Experts, I agree, should have to "step out of the office" (14) and get their boots muddy. I'm also in favour of "relying on evidence" (15), and of being judicious, even cautious, with big ideas. But does this really render our addressing the "big questions" unnecessary? The number of "small steps" we might take in any area is huge. You can't test every possible permutation and combination through RCTs. Something has to guide your choice of questions. Of course you learn from experiments and experience generally, but your thinking guides further experiments. Big ideas help to guide our search for evidence without wasting time and effort on areas unlikely to be fruitful.

Perhaps the authors might protest that we are not yet ready to answer the big questions, and so a hasty answer to them would prematurely rule out worthwhile lines of enquiry, and that's why we need more detailed studies. This could certainly be true in some areas. Turning to education, however, it turns out they do have answers to education's big questions, answers that seem to guide their studies and recommendations. Unfortunately, their answers seem to be the wrong answers, even on the strength of the evidence they themselves give.

The "big questions" in education

The "big questions" in education and development are "whether governments ought to, or know how to, intervene" (73). Aid "optimists" (such as Sachs) are generally "government interventionists", while the aid "pessimists" (such as Easterly) are generally in favour of "laissez-faire" (73). Banerjee and Duflo characterise this debate in terms of "Supply-Demand Wars" (72), and coin the term "supply wallahs" for those in favour of "top-down" (81) government initiatives to supply education. They initially define the "demand wallah" position to mean those who believe that private demand has to precede supply (76), which could still imply government provision, but they then use "demand wallah" to imply also a belief that, at least in part, the market should supply education, emphasizing that "Private schooling is the canonical demand-driven strategy" (83).

In their attempt to adjudicate the so-called Supply-Demand Wars, Banerjee and Duflo begin by pointing out huge empirical problems with the supply wallah position. Yes, the Millennium Development Goals (MDGs) committed nations by 2015 to ensure that every child "will be able to complete a full course of primary schooling" (73). Yes, there has been huge progress towards this goal. But there's not much point getting children into school "if they learn little or nothing once they're there" (74). And that turns out to be the reality of what supply wallahs

have been endorsing: In government schools in India, for instance, "50 percent of teachers…are not in front of a class at a time they should be" (74). High levels of teacher absenteeism and neglect lead to incredibly low standards. Pointing to evidence from the Annual State of Education Report (ASER), a comprehensive survey of education in rural India, they note levels of literacy and numeracy so low as to be "stunning" (75). And depressingly, India is not unique. There is not much to boast about regarding educational quality in any "Third World" (75) public education systems. Whatever *quantity* has been mobilized through supply wallah government interventions, there's been little *educational* benefit. So the first step in their argument elicits empirical evidence *against* the supply wallah position.

So is the demand wallah position the right one, then? The authors set out the position: "Since parents are able to respond to changes in the need for an educated labor force, the best education policy, for the demand wallahs, is no education policy. Make it attractive to invest in business requiring educated labor and there will be a need for an educated labor force, and therefore a pressure to supply it" (77).

Banerjee and Duflo present evidence from around the world on this issue, including the effect of call-centre recruitment on the education of girls in northern India, where parents "discovered that educating girls had economic value, and were happy to invest" (77). The evidence seems prima facie to support the demand wallahs' case (76).

Not so fast, say the authors. There are *theoretical reasons* why the supply wallahs instead must be right. The problem is this: "At the core of the demand wallahs' view is the idea that education is just another form of investment: People invest in education, as they invest in anything else, to make more money—in the form of increased earnings in the future" (77). But the "obvious problem" here "is that parents do the investing and children get the benefits, sometimes much later" (77). The upshot is: "Most parents are in a position of power relative to their children—they decide who goes to school, who stays home or goes out to work, and how their earnings are spent" (78). "'Exactly,' says the supply wallah. 'This is why some parents need a push. A civilized society cannot allow a child's right to a normal childhood and a decent education to be held hostage to a parent's whims or greed'" (78).

That's why you need government to tax citizens to build schools, hire teachers and, if possible, make schooling compulsory, as is done in "rich countries" (78). That's why the supply wallah's position is brought back to the fore, no matter that it has signally failed.

This approach seems to go against all that Banerjee and Duflo espouse at the outset of *Poor Economics*. The position was that empirical evidence should guide our interventions, not answers to the "big questions". But we then find that their

central beliefs lead to a reinterpretation of the empirical evidence that they themselves have supplied.

They try to reinforce the conclusion using evidence from conditional and unconditional cash transfers—but actually they again seem less interested in the evidence than in bolstering their big beliefs. Their conclusions from the evidence appear trivial: "income per se matters for education decisions: Jamal will get less education than John because his parents are poorer, even if the income gains from education are the same for both" (80).

This finding, they argue,

> is important, because if parental income plays such a vital role in determining educational investment, rich children will get more education even if they are not particularly talented, and talented poor children may be deprived of an education. (81)

For the authors, this means that leaving education "purely to the market will not allow every child, wherever she comes from, to be educated according to her ability" (81). Whatever that means, it would seem to be a possible argument against leaving *anything* to the market. They write: "Unless we can fully erase differences in income, public supply-side intervention that makes education cheaper would be necessary to get close to the socially efficient outcome: making sure that every child gets a chance" (81). But clearly maximizing the number of children who "get a chance" doesn't imply wholesale government intervention as espoused by the supply wallahs (and which Banerjee and Duflo seem to be endorsing). Bringing down the cost of education to the poorest, e.g., through targeted vouchers, would be compatible with keeping a market in education.

Having arrived now at the desirability of the supply wallah position, "Top-Down Education Policy" (81), they note: "The question, however, is whether this kind of intervention, *even if it is desirable in principle*, is actually feasible" (81, emphasis added).

This might seem odd. Wasn't the first task of their chapter to give pretty convincing evidence that this kind of intervention *hasn't* been feasible? Having now, however, fathomed that such government intervention after all is necessary, they present other, less damning evidence.

Before turning to this discussion, however, it's worth noting the way they elaborate the question, explicitly to illustrate the demand wallah's point of view: "If parents *do not care about education*, isn't there a risk that such a top-down education drive would just lead to a waste of resources?" (81, emphasis added). Putting aside the slander on poor parents, this badly misrepresents the demand wallah's perspective. The objection to Top-Down Education Policy is not that poor parents

"do not care about education". It's that, for systemic reasons, the apparatus of top-down education works badly. The reasons include the lack of knowledge, incentives and accountability within the public sector, carefully outlined by Easterly (2001) in *The Elusive Quest for Growth*. That's what leads to waste of resources, not parental lack of care.

Banerjee and Duflo (2011) agree that public education is "poor quality" (81), sometimes even "dismal" (83). But—"good news"! (81)—public "schools are still useful" (81). They give evidence from Indonesia (General Suharto "decided to go on a school-building spree" from 1973), Taiwan (which made schooling compulsory in 1968) and Malawi and Kenya to show "There is now a significant body of rigorous evidence testifying to the far-reaching effects of education" (81, 82). Banerjee and Duflo seem to think this evidence significant, but it appears to be of absolutely of no use in answering the big questions concerning government intervention, because there are no comparisons exploring whether or not demand-driven education would have been better. Indeed, the authors seem to acknowledge this: The question we need to ask, they say, is: Could "demand-based approaches", including the "canonical demand-driven strategy" of private schools, "work better?" (83).

Enter private schools

When explicitly setting out the demand wallahs' case, the authors point to a set of empirical predictions that, according to them, the demand wallahs should be prepared to make:

> since parents will start to really care about education, they will also put pressure on teachers to deliver what they need. *If public schools cannot provide quality education, a private-school market will emerge.* Competition in this market…will ensure that parents get the quality of schooling that they need for their children. (77, emphasis added)

And again:

> When the benefits of education become high enough, enrollment will go up, without the state having to push it. People will send their children to private schools that will be set up for them…. (76)

The context makes it clear that they say that these private-school markets should emerge even for the poor, not just for the elite and middle classes. In other words,

Banerjee and Duflo have spelled out important and empirically falsifiable predictions that the demand wallahs should be prepared to make.

Is there any evidence on whether private supply of schooling has emerged in response to demand? Reading the opening section of their education chapter (71-83), readers could be forgiven for thinking that only public education is relevant. In the middle section (83-97), *private* schooling is suddenly revealed to be everywhere. Crucially:

> Even before the education experts gave it the heads-up, many ambitious low-income parents around the world had decided that they had to get their children into private schools, even if they would have to scrimp for it. This has caused the surprising phenomenon of cut-price private schools all over South Asia and Latin America. (83)

Indeed, and there's also a host of evidence from sub-Saharan Africa too. And it's important to note that these markets are *huge*. Evidence for instance shows that a large majority, 65 to 75 percent, of schoolchildren in urban slums attend low-cost private schools in countries in sub-Saharan Africa and India (for a summary, see Tooley 2009). In any case, the authors acknowledge that private education, including low-cost private schools, is burgeoning.

I found it odd that Banerjee and Duflo did not link this newly supplied evidence to their earlier discussion about the demand wallahs' predictions about private education—wouldn't it be relevant to point out that their predictions *have turned out to be true*? Leaving this to one side, the authors then go on to explore the important question concerning the *quality* of education. Recall that the supply wallahs haven't been able to solve this problem, according to the authors. So, "Have private schools cracked the problem of the quality of education?" (2011, 83).

The evidence they give is pretty compelling. Private schools—including those charging as little as $1.50 per month—are better than public schools! The World Absenteeism Survey and ASER, earlier used to show the deficiencies of the public system, also show how private schools, including low-cost private schools, are doing much better. Teachers are absent far less often in private than public schools; achievement is much higher in private than public schools. Banerjee and Duflo also give evidence from the "Learning and Educational Achievement in Pakistan Schools" or LEAPS[2] survey (84). LEAPS shows that in rural Punjab

2. Andrabi et al. (2007) use LEAPS as an acronym for "Learning and Educational Achievements in Punjab Schools". A logo featured in the header on the current LEAPS website (visited June 25, 2012) contains the words "Learning and Educational Achievement in Punjab Schools". The "About the LEAPS project" web page (**link**) speaks of "the Learning and Educational Attainment in Punjab Schools (LEAPS) project". Text on the LEAPS homepage (**link**), however, refers to "The 'Learning and Educational Achievement in Pakistan Schools' project", which is also the name as given by Banerjee and Duflo (2011, 84).

"children in private schools were 1.5 years ahead in English and 2.5 years in math relative to children in public schools". The authors neglect to tell us that the private schools were also 1.5 years ahead in Urdu—in case anyone thought their language advantage might be unfair because private schools were often English-medium (Andrabi et al. 2007, 12). Banerjee and Duflo also don't tell us that the private schools achieved these academic advantages for around half the cost of the public schools, which might also be relevant to an economic discussion of education. Importantly, the authors *do* tell us that the achievement advantage is not just because private schools get children from richer families: "The gap in performance between private- and public-school students was close to ten times the average gap between the children from the highest and lowest socioeconomic categories" (2011, 84).

So they conclude "children in private school learn more than children in public schools" (84). Remember, this evidence is largely about low-cost private schools, serving exactly the kinds of children that development experts say they want to see served.

Isn't this enough to support the demand wallahs against the supply wallahs? Shouldn't Jeffrey Sachs defer to William Easterly?

Rescued by big beliefs

For *Poor Economics*, the answer is no. The authors agree, private schools are better than public. However, we shouldn't be misled into thinking that this strong evidence provides support for a definitive answer to "the supply-demand wars" favouring the non-government route. No, this is not the case because private schools are not "as efficient as they could be" (84).

Now, many things are not "as efficient as they could be", but we still prefer them to the alternatives. Why wouldn't the same be true of private education, and especially those low-cost private schools that everyone agrees are already better serving the poor? Banerjee and Duflo offer two distinctive arguments as to why not:

First, Banerjee and Duflo explain that some "simple interventions" in public schools show how inefficient private schools are (84). One of these interventions was undertaken by Pratham, one of India's largest NGOs, concerned with raising educational standards for the poor. The program, called Balsakhi (literally "children's friend"), took children in public schools who had learning difficulties from each classroom and assigned them to a young woman from the community trained for one week. This programme "generated very large gains in test scores for these children". In some places this was "about twice the magnitude of the average gains from private schooling that were found in India" (85).

This is an odd style of comparison to make from authors who believe in randomised *controlled* trials. Surely if you're comparing public and private, you should not use this one-sided test, where you're comparing public schools *with* special services to private schools *without* such special services?

Perhaps what the authors are trying to do is show that private schools are not as good as they could be, because Pratham came up with a "simple intervention" that gave results up to twice as good as what private schools could do. What they appear to be claiming is that the Pratham programme was cheap and cheerful and easy to implement. On the contrary, the programme needed funding by generous donors, and was based on Pratham's extraordinarily long and fruitful experience in the field. Moreover, just because Pratham's teacher training lasted only one week doesn't mean that it was easy to develop and would be easy to replicate. Indeed, it probably took more resources for Pratham to distil the requirements for good remedial teaching into a one-week course than it would take others with less experience to create a one-year or longer programme.

In other words, not only are the authors not comparing like with like, they are overlooking the fact that the "simple intervention" they're describing might instead be rather complex, based on many years of experience. If that's the case, then their comparison between what low-cost private schools can achieve—run by inexperienced school proprietors with no outside funding or assistance using largely untrained teachers—with what Pratham achieved—with their massed ranks of educational experts and generous donors—seems extraordinarily inappropriate.

Similarly, the authors give examples of remedial reading camps in Bihar, where again Pratham stepped in with one week's training for volunteers and government teachers. The reading gains, they say, were large:

> If volunteer and semi-volunteer teachers can generate such large gains, *private schools can clearly adopt* the same kinds of practices and should do even better. Yet we know that in India a full one-third of fifth-graders in private schools cannot read at first-grade level. Why not? (86, emphasis added)

The argument again seems hardly well thought through. If private schools "clearly" can adopt the same methods, is it because the methods are based on ideas so obvious that anyone can easily replicate them? Again, this suggestion underestimates the sophistication of Pratham's interventions. Or is it because information on Pratham's methods is freely and easily available, so private school managers can easily replicate them? Interestingly, the authors say that similar work is being done in Ghana (86), but although I spend a considerable amount of time working in education in Ghana, and have close contacts with Pratham's leaders in

India, I'd not heard of it before, and I'm an academic with access to large networks. Perhaps the local private-school entrepreneurs living and working in poor areas, not necessarily being connected to the Internet, do not have access to the kinds of information that the authors of *Poor Economics* suppose are so freely available.

It seems to me that Banerjee and Duflo's big beliefs impel their disappointment. They say:

> No doubt, some of the usual reasons that markets do not work as well as they should are at work here. Perhaps there is not enough competitive pressure among private schools, or parents are not sufficiently informed about what they do. (86)

My objection is not to their having big beliefs, but to their holding these beliefs irrespective of where their evidence and argument seem to be leading them.

Finally we come to what Banerjee and Duflo declare as "one key issue" that is "unique to education" (86). This is their second line of argument concerning the inefficiencies of private schools. It is, I believe, a very important and profound argument. But far from supporting the role of government in education for development, as they appear to want to use it, it seems in fact to do the opposite.

The curse of expectations

The crux of this argument is the "peculiar way in which *expectations about what education is supposed to deliver* distort what parents demand, what both public and private schools deliver, and what children achieve—and the colossal waste that ensues" (86, emphasis in original).

The problem they've located they call "The Curse of Expectations" (86), which means that poor parents "see education as a lottery ticket, not as a safe investment" (87). This is partly because the stakes are so high: poor parents "seem to see education primarily as a way for their children to acquire (considerable) wealth" (87).

In most developing countries, the ministry of education within the government either sets or regulates examinations that children must take at the end of a specified period of schooling, and these examinations are the sole gatekeepers that decide if a child can continue on with higher levels of schooling or further education: Without passing these examinations, the better paid and secure government jobs will be out of reach, and in many developing countries it is precisely this kind of employment that poor parents aspire to for their children. But the government imposed exams might be nine, ten or even twelve years after a child

gets enrolled in a school. So how can parents know that their school is helping their children achieve that goal?

It's a huge problem, and one which clearly complicates accountability in the private school market—Banerjee and Duflo are surely correct in that. There *are* informal methods a parent who has chosen a private school can use, which the authors do not acknowledge. Parents talk to each other about how well their children appear to be doing at school. They compare how frequently children's notebooks are marked, and homework given, or how well children speak English. Even though they might not speak English themselves, they can often tell whether or not children are speaking it well amongst themselves.

These things are important. But the parents won't necessarily tell how well the child is going to do in the public examinations. So, as Banerjee and Duflo address the matter, in nine, ten or even twelve years' time, when the child takes the public examinations and fails, there is not much for the parent to do. There's no point in blaming the school, she's paid her dues and paid as much attention as she can to the school, so it can't be the school's fault. Just as I can play the lottery with the same choice of numbers for many years and, if I don't win, blame my choice of numbers rather than the lottery, so the parent blames the child not the school. Provided that the private school gets *some* children through the public exams, the school is absolved of blame.

Banerjee and Duflo are correct in pointing out that, under this system, private schools, especially those serving the poor, don't have to be as efficient as they could be. But if, rather than a distant exam to be taken many years hence, there was more proximate pressure from the government's examination system, then perhaps private schools would try harder and perform better. But that is not the system that currently prevails. The current system allows them at least sometimes to get away with being less efficient than they might otherwise be.

Banerjee and Duflo write: "Parents are not alone in focusing their expectations on success at the graduation exam: The whole education system colludes with them" (89). But what is the "whole education system"? It's the system that sets the curriculum and exams. That is, the *government* education system. The problem lies in the curricular and assessment framework that government has imposed. That, it would seem, is something that Banerjee and Duflo would agree with, but then they seem to miss the point and its crucial implications:

> The curriculum and organization of schools often date back to a colonial past, when schools were meant to train a local elite to be the effective allies of the colonial state, and the goal was to maximize the distance between them and the rest of the populace. Despite the influx of new learners, teachers still start from the premise that their mandate remains

> to prepare the best students for the difficult exams that, in most developing countries, act as a gateway either to the last years of school or to college. (89-90)

So for private schools,

> their entire point is to prepare the best-performing children for some difficult public exam that is the stepping-stone toward greater things, which requires powering ahead and covering a broad syllabus. The fact that most children are getting left behind is unfortunate, but inevitable. (94)

It's important to remind ourselves that at least private schools are doing better at this task than public schools. But they're not good enough. And it is this realisation that again leads the authors to disregard all the evidence mounted about the superiority of private over public education.

Let's spell out carefully how *government policy*, in particular government monopoly over curriculum and examinations, works to undermine the effectiveness of private schools, *especially* low-cost private schools serving the poor. There are two types of pressures exerted on private schools in developing countries in this regard. The first is government regulation. It is normal for private schools to have to follow a government-approved curriculum and their children take government-approved examinations in order to be recognised by the government. In the matter of assessment, typically a developing country will approve its own ministry of education's national curriculum and examinations, plus some examinations set by foreign bodies, for example the International Baccalaureate (set by an organisation based in Switzerland) or International GCSEs (set by British organisations). High-end, expensive, elite private schools generally opt out of the government curriculum and exams and go for the international options. This is too expensive for low-cost private schools, which have to stick with the government curriculum and exams. That is, low-cost private schools, in part because they can't afford any other option, must, if they want to be recognised by the government, follow the government-set curriculum and for their children to take government-set exams.

What if schools decline to be recognised by government? Can't they follow any system they want then? In many countries, being an unrecognised school brings a host of problems including the constant threat of closure, so many private-school proprietors want in the end to become recognised, as soon as they can afford to do so (for a discussion of what this might entail, see Tooley and Dixon 2005). But, probably more importantly, this is also where the second set of pressures is exerted on low-cost private schools. Poor parents recognise that for

them, the only show in town is the government curriculum and examination. The signalling benefit from the official certificate—awarded to children who have passed the government examinations—matters much more to poorer parents than it does to richer parents, who will have other ways, such as extensive networks, to help their children along. Poor children are much more dependent on this government certificate to signify they have passed the government-set examinations, this being the gateway to further schooling, higher education and government jobs. Of course poor parents will exert pressure on private schools to prepare their children for the government examinations. Sometimes private schools can get around the need to be recognised by linking with other private schools that are recognised to allow parents to take their exams in those schools. But whatever method is used, the pressure from poor parents works to ensure that low-cost private schools follow the government curriculum-and-assessment route, rather than try to experiment or innovate with other systems that might overcome the severe problems correctly raised by Banerjee and Duflo.

It's curious to blame the private schools for this, rather than the government framework under which they operate. There is no sense in faulting the market for successfully helping people navigate a baneful framework imposed by government.

Interestingly, Banerjee and Duflo come tantalizingly close to this realisation. Focusing on the "huge waste of talent" that these kind of problems and others bring to education, (2011, 95), they point to one private-sector initiative, run by the giant Indian technology company Infosys, that

> has set up testing centres where people, including those without much formal qualification, can walk in and take a test that focuses on intelligence and analytical skills rather than textbook learning. … This alternative route is a source of hope for those who fell through the gaping holes in the education system. (96-97)

Private-sector Infosys is "doing what the [*public*] education system should have been doing" (96).

Now this is the kind of reform that really could help "reengineer education", as the authors dub their solutions. The only problem is that in many developing countries for private schools to opt for such a route would be illegal until the government endorses it as an option. Banerjee and Duflo skip over any difficulties there might be in getting their preferred solutions applied. But is it really so simple to "re-engineer" government-regulated education?

Re-engineering education?

The authors of *Poor Economics* say that they want to approach solutions to educational problems for the poor without respect to the big questions of whether government should or can intervene in education, instead focusing only on the "small steps" necessary for improvements. As they build up their argument, however, it is obvious that big beliefs play a driving role and lead to big answers. The most sensible conclusion for education based on their evidence would appear to be that private is better than public, and that governments should reform their hegemonic systems of curriculum and school exams. But Banerjee and Duflo seem to disregard the very evidence they present, to arrive at more or less the opposite conclusion—that government intervention is the way forward.

They write: "The good news…is that all the evidence we have strongly suggests that making sure that every child learns the basics well in school is not only possible, it is in fact fairly easy, as long as one focuses on doing exactly that, and nothing else" (97). Just as copying Pratham's methods was easy, so too is getting "*every* child" (emphasis added) to read and write and do arithmetic.

But surely they must realise that they're not the first to think of these kinds of interventions? Experts, some even better qualified than Banerjee and Duflo, have been offering solutions to improve public education for decades. Over the last few decades, billions of dollars have been spent on trying to bring about improvements, and the results are mostly discouraging (Easterly 2006). Looking at their solutions, we can suggest that nothing will be different this time.

One of Banerjee and Duflo's preferred models, for instance, is the "no excuses" charter school in America, like those managed by the Knowledge Is Power Program (KIPP):

> These schools have been shown, in several studies based on comparing those winners and losers of the admission lotteries, to be extremely effective and successful. A study of charter schools in Boston suggests that expanding fourfold the capacity of charter schools and keeping the current demographic profile of students the same would have the potential to erase up to 40 percent of the citywide gap in math test scores between white and black children. (98)

But if charter schools like this are so successful, why are they only a tiny proportion of all schools in America? The authors ignore the problem of vested interest groups, like teacher unions, which actively resist such reforms (see for instance Weber 2010 and Tooley 2012). It's dangerously naïve to invoke the charter-school

model as a possible solution without recognising the political forces ranged against it. Or to use their language, it is not *fairly easy* at all to get these kinds of changes in public education systems, for precisely the reasons given by demand wallahs like Easterly, that the incentives of those with power and influence are not aligned with the interests of the poor (Easterly 2001). Here again Banerjee and Duflo invoke Pratham's work, which brings the "good news" that "it takes relatively little training to be an effective remedial teacher" (2011, 98). Earlier they used the example to damn private schools for not embracing these methods. Shouldn't they at least raise the question of why, if these methods are so easy and obvious, public school systems are also not using them at scale? We noted that private schools, especially low-cost private schools, may not have the resources and educational expertise to be able to develop similar methods. The same excuses cannot apply to public school systems, because the Pratham methods were trialled in public schools; clearly the information and practice is there for public systems to embrace more generally, if they wanted to. If the authors had started to explore why public school systems haven't adopted change that is beneficial to the poor, perhaps they would have realised that there are vested political interests, including again the teacher unions that prevent such beneficial changes from being brought into a public education system.

Similarly, the authors say that information technology can help with education for the poor in developing countries. Nearly a decade ago, they say, they did an experiment, again with Pratham, in government schools in Vadodara, India. Children playing a computer game in pairs for only two hours a week made huge gains in maths. For the authors, "This highlights what is particularly good about the computer as a learning tool: Each child is able to set his or her own pace through the program" (100). For this reader, it highlights how difficult it is to effect change in the public system. For what they are saying is that government and Pratham have been sitting on this result for the best part of a decade, and little has happened to change the great bulk of Indian schools. Shouldn't it be asked: Why not?

In their final section, it seems as though Banerjee and Duflo are about to raise these issues: The message of their suggested proposals, involving "scaling down expectations, focusing on the core competencies, and using technology to complement, or if necessary substitute for teachers, *does not sit well with some education experts*" (100, emphasis added). But rather than invoke the problem of vested interest groups, they promptly let the experts off the hook: "Their reaction is perhaps understandable—we seem to be suggesting a two-tier education system—one for the children of the rich, who will no doubt get taught to the highest standards in expensive private schools, and one for the rest" (100-101). But perhaps what is motivating these educational experts is not high-minded ideals like equality and opportunity, but rather the protection of their own interests. Bringing in schools

that focus on core competencies, like charter schools, should be resisted because it threatens union power. Replacing teachers with technology is to be resisted because it hurts teacher livelihoods (see Moe and Chubb 2009).

Do the solutions offered by Banerjee and Duflo live up to the promise of the book's subtitle, "Radical Rethinking of the Way to Fight Global Poverty"? Hardly.

But a rethinking is suggested by the very evidence they have put forward. As Banerjee and Duflo have shown, the private schools are already serving many of the poor better than the public sector, and at lower cost. Efforts in regulatory reform should seek especially to reduce government influence over curriculum and examinations, to reduce pernicious effects pointed out by the authors. Sound ideas and empirical evidence support an attitude of applauding the private school sector and seeking to liberalise its functioning. Policy reforms are likely to be difficult to achieve in the short term. However, by virtue of a change of outlook among intellectuals, officials, NGOs, philanthropists, and others, progress can still be made to improve quality in private schools, extend access to them, and improve market competitiveness. I've suggested elsewhere the kind of possibilities that could be explored, including creating loan funds, investing in curriculum and assessment (to break the government monopoly discussed above), creating chains of schools, targeted vouchers and ratings schemes (Tooley 2007).

Banerjee and Duflo suggest that there are "three Is"—"ideology, ignorance, and inertia" that "often explain why policies fail and why aid does not have the effect it should" (2011, 16). But they appear guilty of letting their own ideological perspective—in favour of government intervention in education—override the evidence they themselves adduce.

References

Andrabi, Tahir, Jishnu Das, Asim Ijaz Khwaja, Tara Vishwanath, Tristan Zajonc, and The LEAPS Team. 2007. *PAKISTAN: Learning and Educational Achievements in Punjab Schools (LEAPS): Insights to Inform the Education Policy Debate*, Executive Summary. LEAPS Project Publications, February 20. **Link**

Banerjee, Abhijit V., and Esther Duflo. 2011. *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*. New York: PublicAffairs.

Easterly, William. 2001. *The Elusive Quest for Growth: Economists' Adventures and Misadventures in the Tropics*. Cambridge, Mass.: MIT Press.

Easterly, William. 2006. *The White Man's Burden: Why the West's Efforts to Aid the Rest Have Done So Much Ill and So Little Good*. New York: Penguin Press.

Moe, Terry M., and John E. Chubb. 2009. *Liberating Learning: Technology, Politics, and the Future of American Education*. San Francisco: John Wiley & Sons Inc.

Tooley, James. 2007. Educating Amaretch: Private Schools for the Poor and the New Frontier for Investors. *Economic Affairs* 27(2): 37-43.

Tooley, James. 2009. *The Beautiful Tree: A Personal Journey into How the World's Poorest People are Educating Themselves*. New Delhi: Penguin; Washington, D.C.: Cato Institute.

Tooley, James, and Pauline Dixon. 2005. An Inspector Calls: the Regulation of 'Budget' Private Schools in Hyderabad, Andhra Pradesh, India. *International Journal of Educational Development* 25(3): 269-285.

Tooley, James. 2012. *From Village School to Global Brand: Changing the World Through Education*. London: Profile Books.

Weber, Karl, ed. 2010. *Waiting for "Superman": How We Can Save America's Failing Public Schools*. New York: PublicAffairs.

About the Author

James Tooley's first job was as a mathematics public school teacher in Zimbabwe. He is professor of education policy and director of the E. G. West Centre at Newcastle University. His book *The Beautiful Tree* was on best-seller lists in India in 2010, and was the winner of the 2010 Sir Antony Fisher Memorial Prize. It builds on his research on private education for the poor in India, China and Africa, for which he was awarded gold prize in the first International Finance Corporation/Financial Times Private Sector Development Competition. His email is james.tooley@newcastle.ac.uk.

Go to Archive of Comments section
Go to September 2012 issue

EJW
ECON JOURNAL WATCH
Scholarly Comments on Academic Economics

ECON JOURNAL WATCH 9(3)
September 2012: 186-209

Why the Denial? Low-Cost Private Schools in Developing Countries and Their Contributions to Education

Pauline Dixon[1]

LINK TO ABSTRACT

Introduction

A large body of research published since the year 2000 has documented the significant and growing contributions of low-cost private schools in slums and villages around the world. Despite the evidence, there are still some within international aid agencies, governments and academia who do not acknowledge the success and potentialities of low-cost private schooling in developing countries. The failure to appreciate such developments and to admit their significance in relation to government schools is a sort of denial. Many are in denial about what is happening in schooling throughout a large part of the developing world.

In this paper I set out the evidence regarding low-cost private schools in Asia and Africa and look at the debates and arguments put forward by those who champion free government schooling as the principal way forward for the poor. I also relate how some international aid agencies have begun to highlight low-cost private schools and how some philanthropists have put such schools on their agendas.

1. Newcastle University, Newcastle upon Tyne, UK NE1 7RU.

Low-cost private schools in Asia and Africa

According to a recent World Bank "World Development Report", poor parents have been getting a bad deal with regards to government education for their children:

> In many of the poorest countries there are enormous deficits in affordable access. Poor people have less access, lower attainment, and lower quality than those better off. In many countries public sector provision is close to dysfunctional and rife with corruption. The technical quality of instruction and learning outcomes are shockingly low, especially among poor people. (World Bank 2003, 111)

The World Bank report recommends that citizens should be "patient", "because making services work for poor people involves changing not only service delivery arrangements but also public sector institutions" (2003, 1). But many poor parents are not sitting around patiently. Owing to high levels of teacher absenteeism and low teacher effort in government schools (World Bank 2010, 7-8), parents have gone in search of alternative schooling for their children. And they have voted with their feet, abandoning government schools for what they assumed were better quality private ones (Tooley 2009).

Extent of private schooling

Between 1999 and 2003 researchers published findings about India (PROBE 1999; Aggarwal 2000; Tooley and Dixon 2002), Pakistan (Alderman et al. 2001; 2003) and Africa (Rose 2002; 2003), adducing a burgeoning sector of low-cost private schools. For example in Haryana, India, private unrecognised schools were found to be operating in "every locality of the urban centres as well as in rural areas" typically adjacent to a government school (Aggarwal 2000, 20). It was estimated that 50 percent of primary school-aged children in Haryana were being educated in the private sector. Indeed, the choice for parents was no longer whether to send their children to school but to "which type of school" (Aggarwal 2000, 21). In Lahore, Pakistan, it was suggested that around half of children from families earning less than $1 a day attended private schools, even when there was a free government alternative (Alderman et al. 2001). Also, researchers identified a "mushrooming" of private schools in parts of Africa, owing to poor quality in

government schools (Rose 2002; 2003). More recent survey and census data from slums and shanty towns around the world expand upon these initial indications.

Research teams have distinguished two types of fee-paying private schools. First, some private schools function below the radar, operating within the extra-legal sector; these schools are referred to as unrecognised or unregistered. Second, there are those that have gained recognition from the government, i.e., they supposedly abide by regulations and are termed recognised or registered. Typically, unofficial payments, i.e., bribes, are made by school owners to education officials, irrespective of whether the school meets the criteria, to gain recognition (Dixon and Tooley 2005). The government regulations generally do not focus on those inputs that stimulate achievement and quality and are often impossible for the school owners to achieve, e.g., requiring a 4,000 square meter playground or that teacher salaries be equivalent to government ones inflated by the political influence of teacher unions. During an inspection visit, the government official asks the school owner either for monetary payment or a gift in order for him to be able to sign relevant documents. According to one District Education Officer in Hyderabad "The whole system is corrupt, the regulations are flexible, open to bribery and corruption. Bribery is a possibility, they can bribe me too!" (Dixon and Tooley 2005, 46).

Large numbers of private schools have been found in several impoverished urban areas of India. In about 19 square miles of notified slum areas[2] in the three poorest zones of Hyderabad, Andhra Pradesh, at least 65 percent of enrolled children (169,260 from a total of 262,075) are attending private unaided schools. The survey located a total of 918 schools, of which 335 (36.5 percent) were unrecognised private schools, 263 (28.6 percent) private recognised schools, and 320 government schools (35 percent) (Tooley et al. 2007b).[3] According to Padma Sarangapani and Christopher Winch (2010, 504-505), the private schools in Hyderabad's slums are especially valued by Muslim parents, whose community has been neglected and underserved by the Indian state. Surveys carried out elsewhere in India, as well as Africa, show that parents of all faiths value private schools (Rangaraju 2012; Dixon and Tooley 2012; Tooley et al. 2005; Tooley and Dixon 2007; Tooley et al. 2007a; Tooley et al. 2008; Härmä 2011).[4]

---

2. Government of Andhra Pradesh (1997, 40-73).

3. The majority of unrecognised private schools (60.8 per cent) are nursery and primary providers, with around one-third providing all sections. Most (74.3 percent) recognised private schools are 'all through' schools, providing all sections. Three-quarters of government schools were reported to be primary-only schools, with 17.2 percent providing primary and secondary sections only (Tooley et al. 2007b, 546).

4. Several Indian communities mentioned in this paper, including Patna, East Delhi's Shahdara area, and the states of Kerala and Manipur, have sizable religious minority populations. Meghalaya state has a Christian majority. By contrast, Haryana state is dominated by Hindus.

In East Delhi, in a 20 square km slum area called Shahdara, at least 66 percent of schools are private unaided schools. Out of a total of 265 schools located, 73 are unrecognised schools and 102 are recognised unaided private schools (Tooley and Dixon 2007).[5]

TABLE 1. Schools in selected urban areas of India

	Notified slum areas, Hyderabad		Shahdara, East Delhi		Patna	
	No. schools	% Schools	No. schools	% Schools	No. schools	% Schools
Government	320	34.9	71	26.8	336	21.3
Private recognised and aided[6]	49	5.3	19	7.2	14	0.9
Private recognised but unaided	214	23.3	102	38.5	56	3.6
Private unrecognised	335	36.5	73	27.5	1168	74.2
Total	918	100	265	100	1574	100
Total private unaided		60%		66%		78%

The latest research from Patna, Bihar, shows that in all 72 wards of the whole city there are a total of 1,574 schools, of which 79 percent (1,238) are private, compared to 21 percent (336) government. Sixty-nine percent of the 1,224 unaided private schools are low-cost, serving the poor and charging less than Rs. 300 ($5.61[7]) per month. Official data from the District Information System for Education (DISE) does not include unrecognised schools—three-quarters of the schools in the city, enrolling two-thirds of school-going children (Rangaraju et al. 2012).

In rural India, it is estimated, 24 percent of children aged six to fourteen are enrolled in private schools, and in the states of Haryana, Kerala, Manipur, and Meghalaya the share exceeds 40 percent (Pratham 2011, 58). In Pakistan there are around 47,000 private schools, and they cater to about one-third of primary-school enrollees (Andrabi et al. 2007, vi).

Some similar results have been reported from Africa. In the poor urban and peri-urban areas of Lagos state in Nigeria, 75 percent of school children are attending private schools. In the district of Ga, Ghana, which has about 500,000 inhabitants—around 70 percent of whom live on or below the poverty line—75

5. The private unrecognised schools typically cater for primary- and nursery-aged children, with only 1.4 percent providing primary and secondary classes. Some private recognised schools (around 9 percent) cater for all sections, but again the majority are primary schools. Around half of the government schools cater for nursery and primary sections and one quarter are primary-only, with just over 10 percent catering for all sections and the remaining 10 percent primary and secondary classes (Tooley and Dixon 2007, 210).
6. Private aided schools are all recognised and are run by private management, with teacher salaries being paid by the government.
7. The conversion rate as of May 2012 was about $1 to Rs. 53.

percent of the 779 schools located were private, these serving around 65 percent of children (Tooley et al. 2005; 2007a).

Quality and cost of private schooling

Private unaided schools in India predominantly charge monthly fees. Typically there is a statistically significant difference between the fees charged in recognised and unrecognised schools, with the former consistently charging higher than the latter at each class level. The data from schools in Hyderabad show that the mean fees at unrecognised schools for fourth grade would cost about 4.2 percent of the monthly wage for a breadwinner on minimum wage, while a recognised school fee would be around 5.5 percent (Tooley et al. 2007b, 548)[8]. In Africa, private schools predominantly charge term fees. As in the India case, statistically significant differences are found between recognised and unrecognised fees; however, private schools are more expensive for the poor in Africa. For example, in Nigeria fees account for around 12.5 to 13.5 percent of minimum wage for Primary 1 and Primary 4 class (Tooley et al. 2005, 133). But it should be remembered that many private schools provide scholarships as well as reduced fees for those who are struggling, as will be discussed below.

In the slums of East Delhi, teachers in private schools are paid about 40% of what government teachers are paid per pupil (Tooley and Dixon 2007, 212). Private school teachers there had slightly higher rates of absenteeism but showed greater commitment to classroom teaching while working (216). A similar finding regarding commitment was made in Hyderabad; when researchers called unannounced there, more than 90 percent of teachers in private school were teaching, compared to only 75 percent in government schools (Tooley et al. 2007b).

According to Monazza Aslam and Geeta Kingdon (2007), when looking at student attainment in India, teachers' levels of training and certification count for little:

> [M]ost of the standard teacher resumè characteristics (such as certification and training) often used to guide education policy have no bearing on [a] student's standardised mark. (22)

8. Mean fees for first grade in recognised private unaided schools are Rs. 95.60 ($1.80) per month (using $1=Rs. 53/-), compared to Rs. 68.32 ($1.29) per month in the unrecognised schools. At fourth grade, the same figures are Rs. 102.55 ($1.93) compared to Rs. 78.17 ($1.47). Minimum wages for Andhra Pradesh are set in the range from Rs. 25.96 (49¢) to Rs. 78.77 ($1.49) per day (2001 figures, from Government of India 2005), with workers in Hyderabad (who will be non-agricultural) typically at the higher end. A wage of Rs. 78/- ($1.47) per day translates to about Rs. 1,872/- ($35) per month (assuming 24 working days per month).

What seems to have more effect is that:

> Good private schools are…able to retain better teachers by renewing their contracts and firing the less effective ones. (23)

James Tooley (2009, 177-178) reports that, across several mostly urban localities studied in India, Nigeria, and Ghana, facilities were generally better in private than government schools, including access for children to drinking water, electricity, and teaching aids and materials. Private schools also placed fewer children in each class, i.e., they had smaller pupil-teacher ratios (174-175). In China, where private schools are generally established in remote villages not serviced by public schools, facilities were found somewhat superior in government schools, but pupil-teacher ratios were comparable between private and public schools (185-186).

Parents in India also want their children to attend private schools because they purportedly teach all subjects in English. Many, perhaps most, private schools are English medium, while government schools generally teach English as a subject and typically only from the age of 11 years (Tooley and Dixon 2002). Parents believe that learning through English helps their child attain better opportunities after school in both employment and further education. Parents typically are of the opinion that if their child can communicate in English this then provides them a better chance of lifting themselves and their families out of poverty (Sen and Blatchford 2001, Mitra et al. 2003).

As for pupil achievement, what do the data show when comparing government and low-cost private schools? Testing of children in various subjects including maths, English, and home language has been undertaken in developing countries around the world. Data have also been gathered on pupils' family background, innate ability, school and teacher characteristics in order to control not only for school choice and therefore selection bias, but take into account peer influences and other variables that might affect achievement. Various analytical techniques have been used to analyse the data, including multi-level modelling and the Heckman-Lee procedure. Typically the results show that pupils in low-cost private schools outperform those in government schools, and at a fraction of the teacher cost (Tooley et al. 2011; Tooley et al. 2010; French and Kingdon 2010; Pratham 2010; Andrabi et al. 2010; Andrabi et al. 2007).

For example, Rob French and Geeta Kingdon (2010) show that in rural India pupils in private schools significantly outperform government pupils:

> [T]here is consistent evidence of a private schooling advantage throughout the methodologies… and after controlling for age and gender,

> private school attendees have cognitive achievement between 0.20 and 0.25 standard deviations (SD) higher than government school attendees. This is about seven times the effect of gender, and almost equal to the effect of an extra year of education, on average over the age range 6–14. (21, 27)

In Pakistan, children from public schools perform 0.8 to 1.0 standard deviation lower on independently administered tests than do "equivalent children in private schools" (Andrabi et al. 2010). The LEAPS project shows also in Pakistan that children in government schools would need 1.5 to 2.5 years to catch up with children in low-cost private schools (Andrabi et al. 2007, xiv-xv). In Orissa state, India, the private school effect is positive and statistically significant in maths and reading (Goyal 2009).

To summarise: Research in several urban areas of the developing world shows that, in the shanty towns and slums, more children attend low-cost private schools than government ones. Private schools are reported in rural areas, too, and the numbers there are growing. Private schools often have more dedicated teachers, smaller class sizes, and better facilities, even while incurring a fraction of the government schools' teacher costs. And children in low-cost private schools seem to be outperforming those in government schools even after controlling for socioeconomic factors and selection bias.

So why are there some very influential players who refuse to acknowledge the actualities and potentialities of private schools in educating the poor?

The denial

Only government will do

Despite the evidence, there are those who consider government schools as the answer to the need for greater access, regarding private schools as only possibly "filling the gap in poor quality government provision" (Rose 2009, 127). Whatever the cost, and irrespective of what past experience shows, it is imperative that government education systems be *fixed*.

Keith Lewin (2011, 7), director of the University of Sussex development research program CREATE, has recognised that "Growth in the number of [private unaided] schools in some areas has been rapid" in India; however, he believes they "enrol children predominantly from richer households" implying therefore "limits to future growth determined by the affordability of fees to poor

households". Fees, according to Lewin, are restricting the growth of the private sector because poor parents are not able to pay the fees and therefore increase demand which would in turn increase supply. Lewin suggests that it is only increased spending by the Indian government (and any other for that matter) on public secondary schooling that will allow greater schooling access for poor families:

> [P]rivate unaided schools are unlikely to grow to provide secondary education to most outside the top two quintiles of household income. Most growth will therefore be in government or government aided schools. The numbers of local body and government aided schools can grow if there is sufficient capacity amongst suitable stakeholders to take on the responsibilities and meet regulatory and supervision requirements. Where such capacity does not exist government will remain the provider of last resort. (Lewin 2011, 30)

Lewin's belief regarding "richer households" is seemingly contradicted even by an *Oxfam Education Report* from as far back as 2000:

> The private sector is becoming an increasingly important provider of education across much of the developing world. While it plays a smaller role at the primary level than the secondary level, it is growing in importance. The notion that private schools are servicing the needs of a small minority of wealthy parents is misplaced. (Watkins 2000, 229)

Still, that report argues:

> While private schools are filling part of the space left as a result of the collapse of State provision, their potential to facilitate more rapid progress towards universal basic education has been exaggerated. They are unable to address the underlying problems facing poor households, not least because their users must be able to pay…. … The private sector covers a range of providers of variable quality, including many that are of exceptionally low quality. In many countries, only the wealthy can afford good-quality private schools. Private schools of inferior quality are more affordable to the poor, but they do not offer the advantages often assumed for private education. (Watkins 2000, 230)

The author of that Oxfam report, Kevin Watkins, later served as director of UNESCO's EFA ("Education For All") Global Monitoring Report Team. Its reporting has been blunt:

> [L]ow-fee private schools are a symptom of failure in public provision, not a solution to the problem. …*transferring responsibility to communities, parents and private providers is not a substitute for fixing public-sector education systems*. (UNESCO 2008, 131, emphasis in original)

And:

> For the poorest groups, public investment and provision constitute the only viable route to an education that meets basic quality standards. (132)

Research shows, though, that the poor are able to access low-cost private schools that are competitive on quality measures. Daily paid labourers living in slums and shantytowns send their children to affordable, fee-paying private schools, and the poorest groups are able to access private education through scholarships offered by the school owners themselves (Tooley and Dixon 2005; Walford 2011; Ohara 2012).

But leaving things to the haphazardness of the market just won't do, not for UNESCO or Lewin. UNESCO's report states that the growth of the private sector through the market is "unplanned" (seemingly regarding this as a bad thing), and that since "nine out of ten primary school children in developing countries attend public-sector schools, the overwhelming priority should be to improve their standards and accessibility rather than to channel public finance into the private sector" (UNESCO 2008, 164). Planning must remain the order of the day, not bowing to research showing the superiority of "unplanned" schooling.

Allowing human beings to bring all that is good through ownership, entrepreneurship, innovation and competition is all too much for Lewin, as seen in the following somewhat garbled passage:

> Without planning there are the much greater problems of the risks associated with judgements and decisions over influenced by short term political events, populist slogans, causal empiricism, and arbitrary preferences. As a Nigerian colleague puts it "to fail to plan is to plan to fail". (Lewin 2008, 12)

Fixing state education systems must be a priority, for "government failure leads to creeping commercialization through the low-fee private sector", posing "the risk of rising inequity, and the fragmentation of services and standards" (UNESCO 2008, 16).

According to UNESCO there are real limits concerning a market for schooling. Therefore it is crucial to:

> Recognize the **limits to choice and competition**. The development of quasi-markets in education and the rapid emergence of low-fee private providers are not resolving underlying problems in access, equity or quality. While many actors have a role to play in education provision, there is no substitute for a properly financed and effectively managed state education system, especially at primary level. (UNESCO 2008, 21, emphasis in original)

It is only the government that can help the poorest because private schools will not help those who cannot afford the fees, such as orphans or the poorest of the poor:

> [P]rivate (i.e. unsubsidised, for profit) providers will not contribute significantly to achieving EFA ["Education for All"] and the MDGs [Millennium Development Goals]. Private providers will not be the provider of last resort to the poor and will predominantly capture differentiated demand from failing public providers amongst households with relatively high income. (Lewin 2007, 44)

There is a general acceptance that private-school children outperform government-school children. But this is often explained away by the fact that "Private schools are often better resourced than government schools, partly because of the ability of parents to make financial contributions" (UNESCO 2010, 89). Indeed, in private schools parents pay nominal fees. But isn't it true that governments and international aid agencies make large financial contributions to state schools? And as stated above, private schools are outperforming government schools at a fraction of the teacher cost, typically with better facilities, yes, but running on a lower budget and more cost effectively.

A typical statement from those in denial is that these schools only "provide a short term solution" (Woodhead et al. 2012, in press). Even though the data are out there, this is a typical belief concerning the role of low-cost private schools and their contribution to "Education for All" (EFA):

> While private schooling may provide a short-term solution to the educational needs of children in India today, it is unlikely to be the best means of providing education for all children in the longer term in ways that respect equity principles, especially in the absence of strong government regulation including comprehensive public-private partnership arrangements. This is not to say that private schooling does not benefit a large number of children, although knowledge about the extent and nature of those benefits is still relatively weak and the groups of beneficiaries are still somewhat selective by gender, location and poverty

> level. Instead, it is important to emphasise that in so far as it is unable to offer potential benefits to all children, and especially those children who may remain outside of formal schooling or drop-out early, there is little evidence that current growth in the private school sector will make a major positive contribution to the achievement of EFA goals. (Woodhead et al. 2012, 27)

What one has to ask these authors is: Exactly what is the public sector offering? The answer appears to be appalling levels of teacher absenteeism and lethargy, poor quality schools that add little to a child's attainment, and a system that probably made children drop out early because of poor quality and non-existent educational value (Andrabi et al. 2010; Aslam and Kingdon 2007; Tooley et al. 2010).

Denial still exists, even though research has now confirmed that:

- A large proportion of parents have been voting with their feet away from the state sector;
- private schools are outperforming government ones at a fraction of the teacher cost;
- years of funding through national governments and international aid agencies has not improved the standard of state education in developing countries, and much aid is embezzled or never arrives to serve the people it is aiming to benefit (see below); and
- private schools, including those unrecognised and unregistered, are making a significant contribution to education for all targets.

For those still in denial, government schools are the only way forward to educate the poor, even though poor parents and school entrepreneurs do not agree, and strong teacher unions have a stranglehold on government teacher salaries and working conditions. Indeed, one of the reasons to favour the private alternative is that it is less susceptible to the problems and abuses so often concomitant with teacher unions.

What is wrong with acknowledging all the good that is emanating from the low-cost private sector and—instead of trying to fix the unfixable—admiring, praising, and supporting school entrepreneurs and parents?

Inequality and exploitation of the poor

Joseph Chimombo (2009) carried out a case study involving eight schools of four different management types in Malawi, as well as interviews with and gathering data from private provider associations and groups. Like UNESCO, Chimombo comes to the conclusion that fee-paying private schools lead to inequality, favour-

ing "the relatively rich" (2009, 182). When unsuspecting poor parents are lured to the private sector, the warning is that:

> [W]here private schools enrol the relatively poor on a commercial basis, there are risks that the least sophisticated consumers of the service will be exploited and offered poor value for money. (182)

UNESCO (2008) believes there are "acute dangers for equity" (16) with the low-fee private sector, implying "rising inequity" (16). It would seem that UNESCO labours under the mistaken belief that private schools cater only to the more elite in urban areas and therefore marginalise the poor or poorest.[9] Martin Woodhead, Mel Frost, and Zoe James (2012) also suggest that some groups are being marginalised:

> [M]any government schools are becoming "ghettoized" – attended mainly by those from the poorest, most disadvantaged and marginalised groups in society…, which will serve to reinforce wider structural inequalities. (Woodhead et al. 2012, 26)

As stated by many including those who have carried out research in Pakistan, this myth should once and for all be dispelled and put to bed, as "evidence suggests that private schools do not cater only to the urban elite, but are also utilised by the poor" (Alderman et al. 2001; Andrabi et al. 2008). According to Tahir Andrabi, Jishnu Das, and Asim Ijaz Khwaja (2008, 340-342) there is evidence to show that private schools are also bridging the gender gap, even in the rural areas of Pakistan where parents are sending their daughters to low-cost private co-educational schools.

However, if UNESCO and others are concerned that the most disadvantaged are becoming the only children attending government schools then surely an option is not to *fix* the government system for the minority, but to provide the minority with a way out! Targeted vouchers or conditional cash transfers could suffice; such policies are more appropriate than devoting more resources to a whole state education system for children who are being "ghettoized" by it.

9. Very oddly, precisely in the midst of complaints that private schools do not enhance equity, the UNESCO report acknowledges: "In the case of slum areas, as in Nairobi, public schools often simply do not exist" (2008, 16).

Unrecognised schools raise "concerns"

In many low-income areas there are more unrecognised/unregistered private schools than government schools (Tooley and Dixon 2005, Tooley et al. 2007b, Tooley et al. 2007a, Tooley and Dixon 2007).

The UNESCO (2008, 16) report expresses a belief that the existence of unrecognised and unregistered low-cost private schools raises "a different set of concerns". Note the word "concerns". Although UNESCO does acknowledge that low-cost private schools are serving parental demand, the fact that some are operating "outside state auspices" (16)—that is, that they are unrecognised/unregistered and cannot attain the requirements of the on-paper rules and laws (which are typically unattainable in a slum or shantytown and recognition is bought with a bribe; see Dixon 2003 and 2004)—requires "concerns" to be raised.

What are these concerns? Are they that the private school cannot afford the bribe that is required to buy recognition, or that the regulations do not address those inputs that target achievement or stimulate the market? (See Dixon 2003; Dixon and Tooley 2005.) Hardly. UNESCO are "concerned" that the schools are not regulated by the state. Never mind that research shows they are regulated by their consumers, who practice both voice and exit. That it has been shown that children in unrecognised schools are outperforming children in government schools (Tooley et al. 2010, Tooley et al. 2011), and that private unrecognised schools have better facilities with more committed teachers, still does not outweigh the apparent imperative to meet impossible regulations that do not target variables affecting attainment and that in practice are cause for bribes paid to the local district education officer (Dixon 2003; Dixon and Tooley 2005).

Those taking government figures at face value, such as Lewin (2011), believe that in most Indian states the numbers of private unrecognised schools are "not large" (2011, 9). Bihar is one of the states cited by Lewin (2011, 8) as being where there are almost *no* private unrecognised or recognised schools using MHRD statistics. As cited above, the latest survey and census data collected in 2011 in Patna, Bihar, shows a large private unrecognised school market. Yet again, those in denial do not look beyond their comfort zone, which may indeed challenge their own philosophy and selfhood.

The only concern UNESCO should have about unrecognised private schools in India is the threat of closure owing to the Right to Education Act rules and laws. With the closures would come the loss of hundreds of thousands of school places for the poor as well as employment for teachers and school owners. UNESCO should be concerned that these children will be lost to education forever.

Choice

UNESCO, despite the evidence, believes that "international evidence remains patchy" and that private schools for the poor offer "little cause for optimism" (2008, 16). The fact that parents are demanding private schools and choosing them above government schools is regarded as not a "positive choice, but as a negative response to perceived – and usually real – failures of the public system" (16). In fact,

> The rapid growth of low-fee private schools is in large measure a symptom of state failure. Chronic underfinancing, often combined with weak accountability, low levels of responsiveness and poor quality of provision, has led millions of poor households to vote with their feet – and their income – to exit public provision. *This is not a prescription either for equity or for accelerated progress towards EFA.* (UNESCO 2008, 16, emphasis added)

Others also question whether private schools are offering poor parents a real choice (Oketch et al. 2008a and 2010; Oketch et al. 2008b; Oketch and Rollerston 2007). Poor parents are regarded as not making choice decisions comparable to their richer neighbours but as being forced to send their children to private schools owing to the poor conditions and lack of places in government schools—interpreted as no choice.

Moses Oketch and collaborators (2010, 31) believe that it "is because there is inadequate supply of public schooling opportunity in the slums" of Kenya that leads to the poor using private schools—they are crowded out by excess demand for public schools. Oketch et al. describe private schools as being of "poor quality" (23, 24), even though they don't give any evidence to support the claim. They therefore ask "why are poor parents paying for poor quality education when they could be getting fee-free schooling in the state sector?" (Oketch et al. 2010, 24). And they note that in the slums of Kenya nearly half of the pupils "attend poor quality fee-charging private schools, in spite of the existing policy of FPE [free primary education] in Kenya" (Oketch et al. 2010, 24). They purport that poor parents don't have any choice apart from low-cost private education. But even if public schools are entirely unavailable, to say that the array of private alternatives do not provide choice is llke saying that Americans lack choice in churches, since no churches are provided by government. What matters is that private options make people's choices richer than they would otherwise be.

Joanna Härmä (2009) states that in rural villages in Uttar Pradesh the poorest do not have the choice to send their children to low-cost private schools as they

cannot afford them. Härmä finds that private school teachers are paid one-tenth of the government teachers (2009, 155), yet she observes "substantive differences" in teaching activity between government and private schools: Children in low-cost private schools in the villages "were without exception…being taught or working on exercises, while there was virtually no teaching taking place at government schools" (158). Incentives to get children to attend government schools included free uniform for girls, midday meal and free textbooks (157-158). But, even with these incentives, "95% of parents stated that their preferred school type was LFP"—low-fee private (158).

In the rural areas investigated, Härmä finds that half of those interviewed could not afford low-fee private schools and this therefore represents "no choice" for the poor (163). Surely it also means that half of the parents do have a choice, even in rural India, and that international aid or philanthropy could provide the half with "no" choice a means to do so. Yet, according to Härmä, "markets do not deliver universal and socially optimal levels of service delivery," leading to the conclusion that "it is socially desirable to reform the government systems, rather than relying on increased marketization to achieve EFA" (164).

Similar is this statement from UNESCO: "In East Jerusalem, schools attended by Palestinian refugees are overcrowded and under-resourced, forcing many students into private sector provision" (UNESCO 2010, 156). Rather than having chosen low-cost private schools, parents are regarded by those in denial as having no option, being "forced" into private provision.

So choice of private schooling is regarded as "not a choice", "forced choice" or "limited choice" by those in denial. The people who are so concerned about choice favour a system dominated by government institutions funded by taxation. They do not seem to notice the incongruity.

User fees are killer fees

There are those who wish schooling to be offered "free" of cost to all children. Typical are UNESCO and Oxfam, which seem to maintain that user fees should be abolished in all developing countries. Fees impose "a considerable financial burden on poor households" (UNESCO 2010, 16). User fees are "killer fees" that "do not work" and "exclude poor people from the services they need most" (Emmett 2006, 46, 47). Governments around the world need to remove user fees in order to get public services "right":

> Abolishing user fees for primary education and basic health care is one of the most important steps a government can take towards getting the politics of essential services right. (Emmett 2006, 84)

Governments should work with trade unions in order for public sector workers to receive pay and housing that is worthy of their service. Abolishing user fees in primary school is the answer to having more children attend school. Rich governments are to blame for the failure of public education provision because they push private initiatives that "unravel" public ones:

> Too often, rich country governments have contributed to this problem by failing to make good their financial promises to poor countries, or by pushing market-based reforms that unravel public systems, and public responsibilities, still further. (Emmett 2006, 18)

Those still in denial do not understand the importance of fees, and they exhibit a demeaning attitude toward parents. Parents have indicated that by paying fees it makes the service accountable to them. Parents feel they can complain if they are paying fees (Tooley 2009; Tooley et al. 2008), if only because they have abstention as a viable strategy. In turn, school owners can complain if teachers are not performing. Owners, too, have abstention as a viable strategy. Private ownership and user fees make people accountable and make the private schools innovative and efficient.

Abuse of aid

International aid over recent years has typically been misdirected, with billions of dollars wasted, stolen, misappropriated, and some would argue facilitating a dependency culture in aid (Moyo 2009; Easterly 2006; Ayittey 2005). Currently education receives about 12% of all government international aid (bilateral and multilateral). "Basic" education is defined by the Development Assistance Committee (DAC) to include primary education, basic life skills for youth and adults, and early childhood education. Such basic education is reported to receive about 40 percent of the total aid to education (UNESCO 2011, 2). Total aid to education, summarized in Table 2, comprises direct aid to education plus 20 percent of general budget support.

However, according to UNESCO, this isn't enough:

> Despite positive trends, however, aid to basic education remains far too low to ensure that all children are able to go to school. Of the US$5.6 billion in aid to basic education, only around US$3 billion went to the poorest countries. The Education for All Global Monitoring Report estimates that these countries need US$16 billion a year to achieve the EFA

goals by 2015, leaving a deficit of about US$13 billion. (UNESCO 2011, 2)

TABLE 2. Bilateral and multilateral aid to education 2009 (US$ millions)

Country or region (selected examples only)	Total aid to education	Total aid to basic education
Sub-Saharan Africa	4,125	2,027
Nigeria	141	45
Uganda	126	57
Ghana	175	91
Kenya	143	74
Mozambique	295	180
Ethiopia	562	286
Tanzania	342	153
D.R. Congo	192	110
India	776	641
Bilateral (gov-gov), all countries	9,657	3,708
Multilateral (including WB, IMF, UNICEF, etc.)	3,768	1,910
Total	13,424	5,618
Source: Tables 2 and 3 in the UNESCO EFA Global Monitoring Report's "Aid disbursements 2002-2009 tables XLS" spreadsheet **(link)**.		

Those in denial suggest keeping to the same formula, but that is not going to work. Even if funds reach government schools, teachers often are not teaching or even turning up. Government school teachers are impossible to fire owing to strong teacher unions. Relying on planners rather than voluntary private action, and giving government-to-government aid to some of the most corrupt governments, is not going to help children get quality education. Large scale embezzlement, corruption, lack of transparency, poor monitoring and insider dealings all belong to the abuse of aid. For example, according to newspaper reports, Kenyan education ministers have been accused of misappropriating $1.3 million of World Bank and DfID funding provided for education projects. In Kenya over the last four years $17.3 million worth of textbooks have been "lost," allegedly through fraud, theft and destruction (Rayner and Swinford 2011).

Not everyone is in denial

There are some other voices, in addition to those cited in the first section, who recognise realities of government school systems in developing countries. Geoffrey Walford (2011) writes that:

> The obvious, but quite unrealistic, answer is that less economically developed countries should improve their government schools. It is unrealistic simply because most of these countries are swimming in corruption so that a great deal of funding simply does not reach the schools and much of what does is misused. Many developing countries also seem to have entrenched teacher unions that not only protect their members' interests (which is wholly legitimate), but also actually act against the interests of the children who should be being taught. Inspection and accountability are feeble, and bribes are a common feature of authority relationships. Cultures do not change fast – certainly not fast enough for these countries to meet their Millennium Development Goals. (411)

The market for low-cost private schooling has attracted the attention of philanthropists, charities and international aid agencies. Thus far the ideas and funding have been focused on providing access to quality private schools for the poorest and stimulating private school supply. Targeted vouchers are being funded in Pakistan by DfID to allow more children to access low cost private schools. DfID is also supporting projects to stimulate private school supply to girls and other marginalised groups. Some private philanthropy is funding targeted voucher projects in Delhi, India (Absolute Return for Kids (ARK)) and supporting an RTC (randomised controlled trial) to consider the effects that vouchers have on the poorest.

Gradually the private sector is being acknowledged by influential policy shifters in governments, philanthropy, and charity. Much denial still needs to be overcome, but the successes of private ownership and markets in schooling are gradually becoming recognized.

References

Aggarwal, Yash. 2000. *Public and Private Partnership in Primary Education in India: A Study of Unrecognised Schools in Haryana*. New Delhi: National Institute of Educational Planning and Administration.

Alderman, Harold, Peter F. Orazem, and Elizabeth M. Paterno. 2001. School Quality, School Cost and the Public/Private School Choices of Low-Income Households in Pakistan. *Journal of Human Resources* 36(2): 304-326.

Andrabi, Tahir, Natalie Bau, Jishnu Das, and Asim Ijaz Khwaja. 2010. Are Bad Public Schools Public "Bads"? Test Scores and Civic Values in Public and Private Schools. Working paper cited with permission from J. Das.

Andrabi, Tahir, Jishnu Das, and Asim Ijaz Khwaja. 2008. A Dime a Day: The Possibilities and Limits of Private Schooling in Pakistan. *Comparative Education Review* 52(3): 329-356.

Andrabi, Tahir, Jishnu Das, Asim Ijaz Khwaja, Tara Vishwanath, Tristan Zajonc, and The LEAPS Team. 2007. *PAKISTAN: Learning and Educational Achievements in Punjab Schools (LEAPS): Insights to Inform the Education Policy Debate*. Executive Summary. February 20. **Link**

Aslam, Monazza, and Geeta Kingdon. 2007. What Can Teachers Do to Raise Pupil Achievement? *Centre for the Study of African Economies Working Paper Series* 273. University of Oxford Department of Economics (Oxford, UK). **Link**

Ayittey, George B. N. 2005. *Africa Unchained: The Blueprint for Africa's Future*. New York: Palgrave Macmillan.

Chimombo, Joseph. 2009. Expanding Post-Primary Education in Malawi: Are Private Schools the Answer? *Compare: A Journal of Comparative and International Education* 39(2): 167-184.

Dixon, Pauline. 2003. The Regulation of Private Schools Serving Low-Income Families in Hyderabad, India: An Austrian Economic Perspective. Ph.D. diss., Newcastle University.

Dixon, Pauline. 2004. The Regulation of Private Schools Serving Low-Income Families in Hyderabad, India: An Austrian Economic Perspective. *Economic Affairs* 24(4): 31-36. **Link**

Dixon, Pauline, and James Tooley. 2005 The Regulation of Private Schools Serving Low Income Families in Andhra Pradesh, India. *Review of Austrian Economics* 18(1): 29-54.

Dixon, Pauline, Ian Schagen, and Paul Seedhouse. 2011. The Impact of an Intervention on Children's Reading and Spelling Ability in Low-Income Schools in India. *School Effectiveness and School Improvement* 22(4): 461-482. **Link**

Dixon, Pauline, and James Tooley. 2012. A Case Study of Private Schools in Kibera: An Update. *Educational Management Administration & Leadership*, forthcoming. **Link**

Easterly, William. 2006. *The White Man's Burden: Why the West's Efforts to Aid The Rest Have Done So Much Ill And So Little Good*. London: Penguin Press.

Emmett, Bethan. 2006. *In the Public Interest: Health, Education and Water and Sanitation for All*. Oxford: Oxfam International.

French, Rob, and Geeta Kingdon. 2010. The Relative Effectiveness of Private and Government Schools in Rural India: Evidence from ASER Data. *DoQSS Working Papers* No. 10-03. Department of Quantitative Social Science, Institute of Education, University of London (London).

Goyal, Sanjeev. 2009. Inside the House of Learning: The Relative Performance of Public and Private Schools in Orissa. *Education Economics* 17(3): 315-327.

Government of Andhra Pradesh. 1997. *Census of India 1991, series 2, Andhra Pradesh: District Census Handbook Hyderabad*. Hyderabad: Government of Andhra Pradesh.

Government of India, Labour Bureau. 2005. *Statistics: Minimum Wages*. Available online at http://labourbureau.nic.in/wagetab.htm (accessed 10 October 2005). **Link**

Härmä, Joanna. 2009. Can Choice Promote Education for All? Evidence from Growth in Private Primary Schooling in India. *Compare: A Journal of Comparative and International Education* 39(2): 151-165.

Härmä, Joanna. 2011. *Education Sector Support Programme in Nigeria (ESSPIN) Assignment Report: Study of Private Schools in Lagos*. Report Number LG 303. Available from http://www.esspin.org/index.php/resources/reports/lagos (accessed July 2012). **Link**

Kremer, Michael, Nazmul Chaudhury, F. Halsey Rogers, Karthik Muralidharan, and Jeffrey Hammer. 2005. Teacher Absence in India: A Snapshot. *Journal of the European Economic Association* 3(2-3): 658-667.

Lewin, Keith M. 2007. Improving Access, Equity and Transitions in Education: Creating a Research Agenda. *CREATE Pathways to Access Research Monograph* No. 1. June. Consortium for Research on Educational Access, Transitions and Equity, University of Sussex Centre for International Education (Falmer, UK). **Link**

Lewin, Keith M. 2008. Four Decades of Educational Planning: Retrospect and Prospect. Paper presented at *Directions in Educational Planning: A Symposium to Honour the Work of Françoise Caillods*, July 3-4. Paris: International Institute for Educational Planning. **Link**

Lewin, Keith M. 2011. Beyond Universal Access to Elementary Education in India: Is It Achievable at Affordable Costs? *CREATE Pathways to Access*

Research Monograph No. 53. January. Consortium for Research on Educational Access, Transitions and Equity, University of Sussex Centre for International Education (Falmer, UK). **Link**

Mitra, Sugata, James Tooley, Parimala Inamdar, and Pauline Dixon. 2003. Improving English Pronunciation: An Automated Instructional Approach. *Information Technologies and International Development* 1(1): 75-84.

Moyo, Dambisa. 2009. *Dead Aid: Why Aid Is Not Working and How There Is Another Way for Africa*. Harmondsworth, UK: Penguin.

Ngware, Moses W., Moses Oketch, Alex C. Ezeh, and Netsayi Noris Mudege. 2009. Do Household Characteristics Matter in Schooling Decisions in Urban Kenya? *Equal Opportunities International* 28(7): 591-608.

Ohara, Yuki. 2012. Examining the Legitimacy of Unrecognised Low Fee Private Schools in India: Comparing Different Perspectives. *Compare: A Journal of Comparative and International Education* 42(1): 69-90.

Oketch, Moses, Maurice Mutisya, Moses Ngware, and Alex C. Ezeh. 2008a. Why Are There Proportionately More Poor Pupils Enrolled in Non-State Schools in Urban Kenya in Spite of FPE Policy? *APHRC Working Paper* No. 40. African Population and Health Research Center (Nairobi).

Oketch, Moses, Maurice Mutisya, Moses Ngware, and Alex C. Ezeh. 2010. Why Are There Proportionately More Poor Pupils Enrolled in Non-State Schools in Urban Kenya in Spite of FPE Policy? *International Journal of Educational Development* 30(1): 23-32.

Oketch, Moses, Maurice Mutisya, Moses Ngware, Alex C. Ezeh, and Charles Epari. 2008b. Pupil Social Mobility in Urban Kenya. *APHRC Working Paper* No. 38. African Population and Health Research Center (Nairobi).

Oketch, Moses, and Caine Rolleston. 2007. Policies on Free Primary and Secondary Education in East Africa: A Review of the Literature. *CREATE Pathways to Access Research Monograph* No. 10. June. Consortium for Research on Educational Access, Transitions and Equity, Institute of Education, University of London (London). **Link**

Pratham. 2011. *Annual Status of Education Report (Rural) 2010*. Provisional. Mumbai and New Delhi: Pratham Resource Centre. **Link**

PROBE Team. 1999. *Public Report on Basic Education in India*. Oxford and New Delhi: Oxford University Press.

Rangaraju, Bala, James Tooley, and Pauline Dixon. 2012. *The Private School Revolution in Bihar: Findings from a Survey in Urban Patna*. Delhi: India Institute.

Rayner, Gordon, and Steven Swinford. 2011. WikiLeaks Cables: Millions in Overseas Aid to Africa Was Embezzled. *Daily Telegraph*, February 5. **Link**

Rose, Pauline. 2002. Is the Non-State Education Sector Serving the Needs of the Poor? Evidence from East and Southern Africa. Paper prepared for DfID seminar in preparation for 2004 World Development Report.

Rose, Pauline. 2003. From the Washington to the Post-Washington Consensus: The Influence of International Agendas on Education Policy and Practice in Malawi. *Globalisation, Societies and Education* 1(1): 67-86.

Rose, Pauline. 2009. Non-State Provision of Education: Evidence from Africa and Asia. *Compare: A Journal of Comparative and International Education* 39(2): 127-134.

Sarangapani, Padma M., and Christopher Winch. 2010. Tooley, Dixon and Gomathi on Private Education in Hyderabad: A Reply. *Oxford Review of Education* 36(4): 499-515.

Sen, Reena, and Peter Blatchford. 2001. Reading in a Second Language: Factors Associated with Progress in Young Children. *Educational Psychology* 21(2): 189-202.

Tooley, James. 2009. *The Beautiful Tree*. Washington, D.C.: CATO.

Tooley, James, Yan Bao, Pauline Dixon, and John Merrifield. 2011. School Choice and Academic Performance: Some Evidence From Developing Countries. *Journal of School Choice: Research, Theory, and Reform* 5(1): 1-39. **Link**

Tooley, James, and Pauline Dixon. 2002. *Private Schools for the Poor: A Case Study from India*. Reading, UK: CfBT. **Link**

Tooley, James, and Pauline Dixon. 2005. Is There a Conflict Between Commercial Gain and Concern for the Poor? Evidence from Private Schools for the Poor in India and Nigeria. *Economic Affairs* 25(2): 20-26. **Link**

Tooley, James, and Pauline Dixon. 2007. Private Schooling for Low Income Families: A Census and Comparative Survey in East Delhi, India. *International Journal of Educational Development* 27: 205-219.

Tooley, James, Pauline Dixon, and Isaac Amuah. 2007a. Private and Public Schooling in Ga, Ghana: A Census and Comparative Survey. *International Review of Education* 53(3-4): 389-415.

Tooley, James, Pauline Dixon, and S. V. Gomathi. 2007b. Private Schools and the Millennium Development Goal of Universal Primary Education: A Census and Comparative Survey in Hyderabad, India. *Oxford Review of Education* 33(5): 539-560.

Tooley, James, Pauline Dixon, and Olanrewaju Olaniyan. 2005. Private and Public Schooling in Low Income Areas of Lagos State, Nigeria: A Census and Comparative Survey. *International Journal of Educational Research* 43(3): 125-146.

Tooley, James, Pauline Dixon, Yarim Shamsan, and Ian Schagen. 2010. The Relative Quality and Cost-Effectiveness of Private and Public Schools for

Low-Income Families: A Case Study in a Developing Country. *School Effectiveness and School Improvement* 21(2): 117-144. **Link**

Tooley, James, Pauline Dixon, and James Stanfield. 2008. Impact of Free Primary Education in Kenya: A Case Study of Private Schools in Kibera. *Educational Management, Administration & Leadership* 36(4): 449-469. **Link**

UNESCO. 2008. *Overcoming Inequality: Why Governance Matters: EFA Global Monitoring Report 2009*. Paris: UNESCO Publishing. **Link**

UNESCO. 2010. *The Hidden Crisis: Armed Conflict and Education: EFA Global Monitoring Report 2011*. Paris: UNESCO Publishing. **Link**

UNESCO. 2011. Beyond Busan: Strengthening Aid to Improve Education Outcomes. *Education for All Global Monitoring Report Policy Paper* 02. November. Paris: UNESCO Publishing. **Link**

Vasavi, A. R. 2003. Schooling for a New Society? *IDS Bulletin* 34: 72-80.

Walford, Geoffrey. 2011. Low-Fee Private Schools in England and in Less Economically Developed Countries: What Can Be Learnt from a Comparison? *Compare: A Journal of Comparative and International Education* 41(3): 401-413.

Watkins, Kevin. 2000. *The Oxfam Education Report*. Oxford, UK: Oxfam in Great Britain.

Woodhead, Martin, Mel Frost, and Zoe James. 2012. Does Growth in Private Schooling Contribute to Education for All? Evidence from a Longitudinal, Two Cohort Study in Andhra Pradesh, India. *International Journal of Educational Development*, forthcoming.

World Bank. 2003. *World Development Report 2004: Making Services Work for Poor People*. Washington, D.C.: World Bank and Oxford University Press.

World Bank. 2010. *Africa Development Indicators 2010: Silent and Lethal: How Quiet Corruption Undermines Africa's Development Efforts*. Washington, D.C.: World Bank. **Link**

About the Author

Pauline Dixon is Senior Lecturer in International Development and Education at Newcastle University in the North East of England. She is Research Director of the E.G. West Centre at the University and Degree Programme Director of the Masters in International Development and Education. She lectures in economics, education policy and quantitative methods. Dr. Dixon was International Research Coordinator on the John Templeton Project from 2003-2005, the Orient Global Project from 2007-2009 and is currently Research Director on research looking at education in conflict zones. She has around 40 publications, via academic journals, monographs and book chapters. She works as an advisor and external researcher with a number of companies including the English-based international charity Absolute Return for Kids (ARK) in Delhi, India, setting up an education voucher programme as well as introducing improvements in quality to both government and private schools operating in the slum of Shahdara, East Delhi, through the use of synthetic phonics. Her email address is pauline.dixon@newcastle.ac.uk.

Go to Archive of Comments section
Go to September 2012 issue

EJW
ECON JOURNAL WATCH
Scholarly Comments on Academic Economics

ECON JOURNAL WATCH 9(3)
September 2012: 210-233

Was Occupational Licensing Good for Minorities? A Critique of Marc Law and Mindy Marks

Daniel B. Klein[1], Benjamin Powell[2], and Evgeny S. Vorotnikov[3]

LINK TO ABSTRACT

In 2009, the *Journal of Law and Economics* published an article by Marc T. Law and Mindy S. Marks entitled, "Effects of Occupational Licensing Laws on Minorities: Evidence from the Progressive Era." The authors use "Progressive Era" broadly—their data ranges mainly from 1880 to 1940. In their abstract, Law and Marks say: "This paper investigates the effect of occupational licensing regulation on the representation of minority workers in a range of skilled and semi-skilled occupations," and, "We find that licensing laws seldom harmed minority workers. In fact, licensing often helped minorities, particularly in occupations for which information about worker quality was difficult to ascertain" (351). They conclude their article by addressing current policy: "Given that minorities are still underrepresented in many skilled occupations, this suggests that licensing may have an important role to play in helping talented minority workers signal quality" (364).

At the beginning of the paper, Law and Marks quote Walter E. Williams, an outspoken critic of occupational licensing. The quotation represents the viewpoint questioned by Law and Marks; it reads: "Occupational licensing coupled with white-dominated craft unions has been a particularly effective tool for reducing employment for Negroes" (Williams 1982, 90-91; quoted by Law and Marks 2009, 351). As Law and Marks note, Williams belongs to a tradition that includes Reuben

1. George Mason University, Fairfax, VA 22030.
2. Suffolk University, Boston, MA 02108.
3. University of Minnesota, St. Paul, MN 55108.

Acknowledgments*: For valuable feedback we are grateful to Niclas Berggren, David Bernstein, Tyler Cowen, Mike Ford, Daniel Houser, Morris Kleiner, John Majewski, David Skarbek, Shirley Svorny, and Walter E. Williams.

*Omitted upon this article's initial publication due to typesetting error.

Kessel (1958 and 1970), Armen Alchian and Kessel (1962), H. E. Frech (1975), and A. L. Sorkin (1977).[4]

Law and Marks use a state-based difference-in-differences approach to test the effect of licensing on the representation of blacks and women in 11 occupations: accountants, barbers, beauticians, engineers, midwives, pharmacists, plumbers, practical nurses, registered nurses, physicians, and teachers. In their conclusion, Law and Marks write: "Contrary to received wisdom, our empirical analysis suggests that the introduction of licensing did not generally harm black and female workers" (2009, 364).

Such efforts can certainly inform and improve understanding of the consequences of occupational licensing. After scrutinizing the article, however, we find many problems, and some of the problems seem to be quite important. Also, we apply falsification tests to their findings and the results are damaging. All told, the problems suggest that Law and Marks have little basis for upsetting the scholarly tradition indicating that occupational licensing has been detrimental to minorities.

The present critique makes many points. Although the single most important point comes first, thereafter we present our points in an order that facilitates explanation of the details of the Law and Marks article. Some highly important points do not come until later.

The reader may be interested to know that we submitted this paper to the *Journal of Law and Economics* in a form essentially identical to the present one, but it was rejected.

Problems with their data

A Census-reported practitioner in a licensing state is not necessarily licensed

Licensing is instituted principally at the state level.[5] States imposed licensing on, say, plumbers at different dates. The difference-in-differences approach aims to detect the effects of the imposition relative to what is happening in the states without the law. In principle, this method helps to control for trends occurring apart from licensing, such as overall movements of blacks or women into non-

4. Kleiner (2006) and Stephenson and Wendt (2009, 184-189) summarize the economic literature on licensing.

5. Freeman (1980, 167) offers the following useful clarification: "state laws differ in applicability, some requiring licensing in cities above a certain size and others permitting cities to license occupations."

agricultural work. For data on when a state imposed licensing on an occupation, Law and Marks draw from the report entitled *Occupational Licensing Legislation in the States*, produced and published by the Council of State Governments (1952).

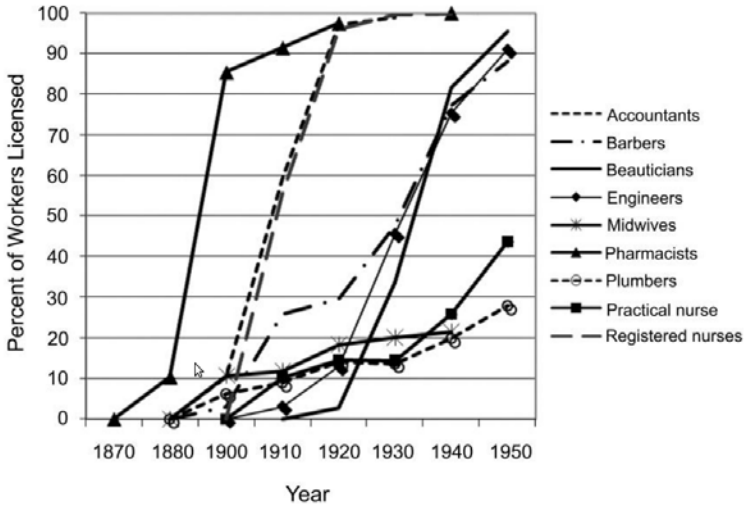
For data on the number of practitioners in an occupation, such as plumbing, Law and Marks do *not* draw from records of plumbers licensed. Instead, their data on occupational employment comes from the Integrated Public Use Microdata Samples (IPUMS-USA 2010) of the Census of the Population. As they note in their paper (2009, 356), the data on occupation is self-reported.

People sometimes practice without a license. The Census was not and is not used to enforce against unlicensed practice, so there would be seem to be little reason for an unlicensed plumber or nurse not to report his or her true occupation. Law and Marks propose that, when representation of blacks and women increased, it was *because* licensing introduced a way for blacks and women to assure quality. Yet we do not know that those blacks and women had a license.

That a Census-reported practitioner in a licensing state is not necessarily licensed is a fact that should greatly affect the researcher's treatment of the data and results. Remarkably, however, Law and Marks make no mention of this fact, and thus give no discussion of how it affects an understanding of the data and results. It is curious that, in a *Journal of Economic History* article by Law and Sukkoo Kim (2005), an article that also uses the Census data, Law and Kim suggested that one possible reason that licensing did not appear to reduce entry was "because enforcement of early licensing laws was weak" (740; see also 731 n. 23). Yet the importance of unlicensed practice is never acknowledged in the 2009 article by Law and Marks.

In their article, Law and Marks make almost no statement to the effect that they are treating Census-reported practice in a licensing state as licensed practice. There is one explicit statement that would be true only if all Census-reported practice in a licensing state were licensed practice. Here we reproduce their Figure 1 (Law and Marks 2009, 355), which charts occupations over time. The vertical axis is labeled "Percent of Workers Licensed." The line for plumbers, for example, leads the reader to think that, by 1950, 28 percent of all plumbers in the country were licensed. What their data actually tell us is that 28 percent of plumbers in 1950 *worked in licensing states*. In light of the fact that not all Census-reported practice in a licensing state is licensed practice, we see that the axis label "Percent of Workers Licensed" is incorrectly labeled. The correct label would be "Percent Working in Licensing States."

Figure 1. A direct reproduction of Law and Marks's Figure 1 (2009, 355): The label on the vertical axis would be sound only if all Census-reported practice in a licensing state were licensed practice.



With that one exception, Law and Marks avoid ever saying explicitly that all Census-reported practice in a licensing state is licensed practice. Their statements throughout the paper, about the impact of licensing laws on minority representation in the occupation, let the reader fall into thinking that all Census-reported practice in a licensing state is licensed practice. In never pointing out the falsehood of such a thought, Law and Marks suggest *implicitly* that all Census-reported practice is licensed practice.

Law and Marks might defend themselves by saying that they were careless in labeling the vertical axis of Figure 1, and, further they might claim that we can *suppose* that, in a licensing state, the ratio of licensed practice to Census-reported practice is generally the same for the majority (white/men) as it is for the minority (blacks/women). That is, they might assert that the fact that not all Census-reported practice in licensing state is licensed practice merely introduces classical measurement error. One cannot find a defense of such a supposition in Law and Marks (2009), for they do not even acknowledge the issue.

But there are reasons to suspect, rather, that in a licensing state the ratio of licensed practice to Census-reported practice would be higher for whites/men than it would be for blacks/women. Let's examine plumbers as a concrete example.

In his book *The Negro and Organized Labor* (1965), Ray Marshall explains that in 1953 the Maryland Senate conducted an investigation of the Maryland's State Board of Plumbing Commissioners to "explain evidence of discrimination against plumbers in the state" (Marshall 1965, 115).[6] It disclosed that there were 3,200

licensed plumbers in Maryland in 1953, and only two were black (115-116). Meanwhile, the Census for Maryland for 1950 showed 6,265 plumbers, of which 6,169 were white and 96 were black.

TABLE 1. Plumbers in Maryland, ca. 1950: whites and blacks

	A	B	C	D
Maryland plumbers	1953 Maryland Senate investigation: LICENSED	1950 Census-reported plumbers	A as % of B	B as % of A
White	3198	6169	51.8%	192.9%
Black	2	96	2.1%	4800.0%
ALL	3200	6265	51.1%	195.8%
Sources: Column A: Marshall (1965, 115-116); column B: U.S. Census (IPUMS-USA 2010).				

In Table 1, the numbers in column A are the numbers from the 1953 Maryland Senate investigation, and the numbers in column B are as we find them from the 1950 Census. Column C shows that 51.8 percent of white plumbers were licensed while only 2.1 percent of black plumbers were licensed. Table 1 constitutes simple, straightforward evidence that the licensed percentage among whites was much greater than the licensed percentage among blacks. If one were to suppose—as does Law and Marks's label on their Figure 1—that Census-reported practice were taken as licensed practice, then, as shown in column D, for Maryland in 1950 the number of licensed black plumbers would be overstated by 4,700 percent!

To our knowledge, the information from the 1953 Maryland Senate investigation on licensing and plumbing is the only such information available on actual numbers of licensed practitioners; no one seems to have a data series of licensed practitioners for the period. The Maryland case might be an extreme one. First, the motivation for the 1953 investigation was a concern about the discriminatory effect of licensing in plumbing in Maryland.[7] We can also imagine why unlicensed practice may be much more widespread in plumbing than in several other occupations treated by Law and Marks.

Although the Maryland plumbing data might be an extreme case, it nonetheless illustrates the broad truth that, typically, unlicensed practice was (and may

6. In our research for the present critique, we consulted with David Bernstein, who has published research on how licensing affected blacks in the time period covered by Law and Marks (Bernstein 1994; 2001, ch. 2). Bernstein put us on to this information that bears directly on the issue at hand.

7. Northrup (1944, 24) remarked: "At present [1944] six Negroes are studying nights in the hope of being Maryland's first licensed plumbers of their race." Marshall (1965, 115) reported that these black plumbers had been trying to comply with the state licensing regulations and secure their licenses since 1941, passing a journeyman test only after legislative investigation and adverse publicity in 1949. Marshall further noted that: "[N]o black plumbers passed the examination to become a master plumber."

well remain) much more common among black practitioners than among white practitioners, and probably also for women than for men. For example, Marshall writes: "*as with the electricians*, licensing laws have been used to bar Negroes from the plumbing trade," (1965, 115; italics added), and he relates that in 1962 there were just two licensed black electricians in Atlanta (112).

David Bernstein (2001, 35-36) explicitly warns researchers against using the Census data in just the way that Law and Marks do. He illustrates the hazard with the example from Maryland in 1953, but also he provides ample other illustration, including the following:

> The first African American passed a plumbers' licensing exam in Colorado in 1950 only after pressure from civil rights authorities. There was only one licensed African American plumber in Charlotte in 1968. As late as 1972, Montgomery County, Alabama, had only one licensed African American plumber, and he was able to get his license only after a ferocious struggle with the local plumbers' union. By the early 1970s there were still very few licensed African American plumbers in the United States. (Bernstein 2001, 36)

Bernstein (2001, 38-39) quotes the black barber Ben Taylor protesting to the Virginia state legislature that his colleagues would "dye old Virginia with their blood" before complying with a proposed licensing law on barbers. Stuart Dorsey (1983) discussed reasons why a significant number of black barbers operated without licenses in St. Louis and Chicago. Unlicensed minorities often practice in segmented markets, as has often been observed in taxi markets (Suzuki 1985 and 1995). Segmented markets were especially likely in the pre-Civil Rights South, when state regulation consciously sought to limit public interaction of blacks and whites.

During the decades treated by Law and Marks, fraternal and mutual-aid societies were pervasive (Beito 2000, 2). Membership rates were high among blacks, immigrants, and poorer people. The organizations often provided services that the Census taker would record as nursing, midwifing, doctoring, or teaching. David Beito studied a few black hospitals in Mississippi and writes: "By 1960… [t]he commission gradually shifted from its previous policy of regulatory laissez-faire and started to issue citations" (2000, 195). In other words, enforcement was lax prior to 1960. What was being enforced? The citations were "finding the hospitals guilty of infractions such as inadequate storage and bed space, failure to install doors that could swing in either direction, and *excessive reliance on uncertified personnel*" (195, italics added). The example of the mutual aid societies helps us to realize that many occupations were ripe for unlicensed practice by blacks and by women.

Law and Marks might agree that rates of unlicensed practice were significantly higher among whites/men than among blacks/women, but say that what they are measuring is practice *per se*, regardless of whether the practice was conformant to or in violation of licensing laws. The problem with such a response would be, however, that it goes against the overarching interpretation they give to their findings. That interpretation is that licensing laws provided minorities with a means of signaling their quality, and that *for that reason* their representation in the occupation only rarely was found to decline and in some cases even go up.

Although no one has extensive data about the relative rates at which whites vs. blacks (as well as, men vs. women) were licensed when practicing in a licensing state, we regard a supposition that those rates were roughly equal to be so at odds with scholarly knowledge that it must bear the burden of proof. Law and Marks have done nothing to overcome the burden of proof in justifying the supposition.

It would be reasonable to judge the points of this section as fatal to Law and Marks's article. But their article suffers also from many additional problems.

Is their data on regulation reliable?

The monograph entitled *Occupational Licensing Legislation in the States* (1952) (hereafter *OLLS*) warns researchers to be extremely careful using its data: "[Survey] procedure often results in differing interpretations in responding to particular questions. As far as possible these have been reconciled, but some conflicts doubtless remain, *particularly with respect to dates of initial licensing acts*" (*OLLS*, 9, italics added).

In an attempt to verify the data provided by *OLLS*, we examined the case of the legal profession because in that case the necessary regulatory data was available in two alternative sources. When the regulatory data in the legal profession provided by *OLLS* was compared with regulatory data provided by H. L. Wilgus (1908) and Richard Abel (1989), we found significant discrepancies in when regulations were introduced. *OLLS* reported data on when regulations were introduced in the legal industry for 29 states. For 16 states, *OLLS* reported that the profession did not have licensing regulations, yet Wilgus (1908, 682) reported that those 16 states already did have licensing regulations, and for 12 out of 16 states the discrepancy between the reported years exceeded a decade. Abel's (1989) discussion of the development of licensing regulation in the legal profession supports Wilgus rather than *OLLS*. If Wilgus and Abel are correct, then the data in 55 percent of the states reported in *OLLS* for the legal profession was incorrect. Such a major discrepancy raises doubts regarding the accuracy of the regulatory data on other occupations used by Law and Marks.

Problems with their analysis

Law and Marks examine 11 occupations. For nine—accountants, barbers, beauticians, engineers, midwives, pharmacists, plumbers, practical nurses, and registered nurses—they simply observe when a state licensed the occupation without any regard for the details of the licensing requirements. For two occupations, physicians and teachers, they looked at the effect of two particular educational requirements. Tables 1 and 2 summarize their findings for women and blacks in the eleven occupations. Blank cells indicate that data limitations prevented them from analyzing the occupation while "No Finding" indicates that data was analyzed but that there was not a statistically significant result.

Consider the simpler analysis of nine occupations. For blacks, we see "Positive" for practical nurses and "Negative" for barbers. For women, we see "Positive" for engineers, pharmacists, plumbers, and registered nurses, and no "Negative."

As for the investigation of physicians, they find that the requirement of some pre-med college education helped blacks; the three other effects were statistically insignificant. In the investigation of teaching license requirements, they find that both educational requirements helped blacks, and that both hurt women.

Our Tables 2 and 3 represent the essence of the empirical evidence offered by Law and Marks. Taking the results shown in Tables 2 and 3 at the face value, it would not be surprising if a reader figured, as Law and Marks say, that the evidence indicates that licensing did *not* tend to harm blacks or women; in fact, on balance it seems to have helped in a few instances.

TABLE 2. Summary of the difference-in-differences results of Law and Marks (2009) for nine occupations

	Black	Female
Accountant		No Finding
Barber	Negative	
Beautician	No Finding	
Engineer		Positive
Midwife	No Finding	
Pharmacist		Positive
Plumber	No Finding	Positive
Practical Nurse	Positive	No Finding
Registered Nurse	No Finding	Positive

TABLE 3. Summary of the difference-in-differences results of Law and Marks (2009) for specific licensing requirements in two occupations

	Black	Female
Physician		
4-Year Medical Degree	No Finding	No Finding
Some Pre-Med College Ed	Positive	No Finding
Teacher		
High School Degree	Positive	Negative
Some College Ed	Positive	Negative

In Tables 2 and 3 the word "Positive" appears eight times and the word "Negative" three times. We proceed to point out further problems with Law and Marks's analysis. These problems indicate that there should likely be more Negatives and fewer Positives.

Criteria for including occupations: Sample selection bias

On what basis were the 11 occupations selected? Law and Marks state the criteria as follows:

> For an occupation to be included in our sample, it had to meet three criteria. First, the adoption of licensing regulation had to span at least 2 decades. Second, the occupation had to have a sufficiently large sample in the Integrated Public Use Microdata Samples (IPUMS) of the Census of Population. Finally, at least 1 percent of the people in the occupation had to be either black or female. (Law and Marks 2009, 355)

The final criterion, that at least one percent be black or female—by which they mean one percent in each case (not combined)—could well have selected certain occupations out of the study *because* licensing was so strongly discriminatory. If, for example, attorneys were selected out of the study by the one-percent criterion, it could be that it was because licensing helped to make the field so devoid of blacks and women. The occupation selection method would be biased if it eliminated from the investigation occupations for which licensing kept representation of blacks and women below the cutoff level of one percent.

Criteria for including occupations: In several cases, the participation rates are small

Again, one criterion for including an occupation in the study was that "at least 1 percent of the people in the occupation had to be either black or female" (355). When we look at the Positive findings, the minorities' participation rates are often very low. Reading the descriptive table (Law and Marks 2009, 357), we see that women made up only 1.17 percent of engineers, 1.2 percent of plumbers, and 3.62 percent of pharmacists, while blacks made up only 1.4 percent of physicians. Thus three are within 0.4 percentage points from being excluded from the study.

Law and Marks note, "In those cases for which the number of minority workers is very small, we must be cautious because many state-year cells contain no minority workers in a given occupations" (358). They add: "However, because we have information on many occupations for which different states enacted licensing laws at different times, finding similar results across different occupations gives us more confidence in our analysis" (358). This last sentence seems to say that, because several of the small-participation cases come out Positive, the effect seems to be real. But it could be that these small-participation cases come out Positive because of the problems in their analysis. In that case, Law and Marks use spurious findings to justify one another.

Licensing misidentified: State certification is not licensing

One of the biggest problems with their paper is that Law and Marks lump certification in with licensure. In some of the occupations included by Law and Marks, the government offers the service of certifying practitioners, but does not threaten those who practice without a certificate; they are prevented merely from *using the title* conferred by the certificate. Thus a certification may help signal quality to overcome discrimination, but, unlike a license, it does not create a barrier capable of excluding minorities from a profession.

In Table 3 we see that Law and Marks report four statistically significant results for teachers. But they have misidentified teacher certification as teacher licensing: We believe that in no states were people required to have a license to teach in a private school. There is also the problem of cases where no significant results were found. Law and Marks found no statistically significant relationship between "licensing" and female participation in the accounting profession. Yet accountants had (and have) the option of being certified, becoming a CPA, but are not required to do so to practice. Hence the finding of no adverse impact should be related to certification, not to licensing. This point is also relevant for engineers and for registered and practical nurses. In these occupations Law and Marks claim

that women and blacks benefited from licensure, yet the initial regulation in most states consisted only of certification (*OLLS*, 24). With no barrier to practicing without a certificate, the laws could not hinder minorities from participating in these occupations. It is odd that Law and Marks would use these findings as evidence to support the official rationale for licensing.

OLLS explains the distinction between "optional certification" and "compulsory licensing" (24), but then proceeds to use "licensing" sometimes in the narrow sense that distinguishes it from certification and other times in a broad sense that includes certification. Table 1 in *OLLS* is entitled "Occupations Licensed by States and Dates of Licensing Statutes." By comparing the information in the Table to verbal descriptions in the text (at 25 and 64-77), we see clearly that the Table gives the year *of the state's strongest regulation, even if that is mere certification*. Yet Law and Marks proceed to treat the data as though it is all bona fide licensing.

It is particularly troubling that Law and Marks treat nursing certification as if it is licensure because the move from certification to licensure is the topic of the investigation undertaken by Law and Marks in a new working paper (2012). There, they explicitly state: "Regulation of the two nursing professions has evolved in stages, diffusing gradually across states and also becoming stricter over time. … By 1950 for registered nurses and 1960 for practical nurses, the diffusion of *certification (i.e. voluntary licensing)* was complete. During the subsequent decades, regulation of both nursing professions moved towards *mandatory licensure*" (Law and Marks 2012, 6, italics added). Yet in their 2009 paper they claim to find that "licensing" in the Progressive Era helped black practical nurses and female registered nurses despite the fact that their other research admits that they know nursing was not licensed during the Progressive Era.

No measure of restrictiveness

For the nine occupations listed in Table 2, even when they properly identify licensing, Law and Marks measure only whether an occupation is licensed or not. The degree of restrictiveness of the licenses for each of these occupations varies from state to state and over time in the same state.

For instance, Law and Marks treat a barbering license that requires a three-year apprenticeship plus hours of schooling in the same way that they treat a barbering license that requires only a one-year apprenticeship. In the case of beauticians, Law and Marks treat a license that requires no minimum education and only 1,000 hours of specialized training the same way as they treat the license that requires four years of high school and 2,500 hours of specialized training (*OLLS*, 66). Large differences in licensing fees and educational requirements certainly have

the potential to adversely affect poor minorities. Yet the data Law and Marks use has no ability to control for restrictiveness.

The Law and Marks analysis of physicians looks at whether requiring a four-year medical degree or some pre-med education impacted minorities—mostly arriving at no finding. But physician licenses vary on many margins, not just the two that Law and Marks analyze.[8] It is possible that, on net, licensing had the opposite impact of that shown by Law and Marks, and they just happened to pick one margin of the license that had a different effect.

Most teachers were employed by government

Two of the eight positive findings in Tables 2 and 3 are for black teachers. One problem with these two positives is the misidentification of certification as licensing. But, moreover, according to the National Center for Education Statistics, of the 1919-1920 instructional staff working in either public or Catholic K-12 schools, 93 percent worked at the public schools.[9]

In his study of licensure's impact on blacks, Richard B. Freeman (1980) controlled for government employment. Questions about the direction of causality are especially relevant when relating two variables that are both forms of government policy. Also, differences between the market for school teachers and private markets for other types of practitioners are large. Furthermore, the notion that teaching quality has been improved by certification requirements has been challenged by many scholars (e.g., Lieberman 2007, ch. 4; Hanushek 2011; Goldhaber and Brewer 2000, 141).

We are not arguing that it is illegitimate to study teachers because government is the main employer. Our claim is that because government is the main employer it is illegitimate to use teachers to speak about the effect of licensing on minorities more generally, as Law and Marks do.

8. Vorotnikov (2011) does a panel study of licensing to practice law, examining impact on incomes and service quality. He controls for five different license requirements. Three requirements increased incomes and two had no statistically significant effects. One requirement increased quality and four decreased quality. It would have been misleading to just pick any single one of these requirements and claim it was the effect of licensing.

9. The National Center for Education Statistics shows that in 1919-20 instructional staff was 678,000 at the public schools and 49,516 at Catholic schools (Snyder and Dillow 2011).

Registered nursing: Licensing helped women overcome discrimination?

One of the eight positives is for women registered nurses. In one of Law and Marks's descriptive tables (2009, 357) we see that, covering the years 1900-1940, women accounted for 97.3 percent of all registered nurses. And yet Law and Marks interpret the finding on registered nurses as evidence in favor of their theory that licensing helped to advance the "minority." On their interpretation, licensing of registered nurses provided women a new and special assurance device, enabling them to overcome discrimination—that is, discrimination in favor of male nurses. Law and Marks do not pause to address the incongruity of their ideas with the findings they invoke.

When Law and Marks come to women in teaching, however, where women made up 78.7 percent (357) and where the finding is *negative*, they *do* pause to notice the gender proportions: "Taken at face value, the results for women are consistent with the standard hypothesis that argues that entry barriers facilitate discrimination. However, we are uncertain as to whether this is the correct interpretation, since in occupations that are disproportionately female, it is unclear which sex is the target of discrimination" (363). Now, for a finding that goes against their view, they bring up a reason to discount it. But that same reason went unmentioned for a finding that supports their view. They cannot have it both ways.

Failure to control for craft/trade unions

According to Leo Troy (1965, 2), in 1920-21 union membership as a percent of nonagricultural employment was about 19.5 percent. Craft unions were often like guilds serving customers in the marketplace, as opposed to workers seeking collective bargaining. As is well known from standard works such as Sterling Spero and Abram Harris (1931) and many by W. E. B. Du Bois, such unions often played a major role in limiting blacks' access to occupations (Higgs 1980, 85-86). Law and Marks note that "Unequal access to education, union control of entry in certain trades, and imperfect credit markets greatly restricted minority entry into certain high-skilled occupations" (2009, 353).[10] In fact, their lead quote in the article notes "Occupational licensing *coupled with white-dominated craft unions* has been a particularly

10. The sentence preceding this quote reads, "Because licensing laws were introduced in the Progressive Era, it is important to interpret our findings in light of the historical facts about labor markets for minority workers" (Law and Marks 2009, 353). However, Law and Marks do not seem to let this historical interpretation prevent them from making general claims about licensure and giving policy advice today.

effective tool for reducing employment for Negroes" (Williams 1982, 90-91; italics added).

Many, if not most, craft unions had a legacy of discrimination. Law and Marks mention unions only in passing, as quoted above. In the statistical analysis they do not control for the influence of unions. Craft unions were not prevalent among accountants, midwives, physicians, and practical nurses. Yet union membership was prevalent in some occupations where Law and Marks find positive impacts, namely, engineers, pharmacists, plumbers, registered nurses, and teachers.[11] And they were prevalent for beauticians, for which Law and Marks arrived at no finding, and for barbers, for which they found a negative impact.

From the point of view of those seeking to exclude minorities, licensing may have served as a substitute for craft unions. We can imagine that a positive finding by Law and Marks comes not because licensing helps minorities to signal quality, but because licensing was often pursued by exclusionists because minorities had not been excluded by unions. In other words, where unions had been effective, licensing was less keenly pursued by exclusionists. On the conjecture, Law and Marks could find a positive effect of licensing only because they fail to control for the negative effect on minorities that unions were having in states that didn't adopt licensing.

No control for changes in the minority population

The decades treated by Law and Marks was a period of great interstate migration, particularly by blacks leaving the South. Suppose that states like Illinois that had high black in-migration tended to be states that imposed licensing. It is possible that, after licensing, in-migration increased minority practice in spite of licensure. When testing factors that may have influenced the adoption of licensing, Law and Marks include black/woman share of the labor force (356). But in the difference-in-differences investigation itself they do not have a variable that would control for interstate migration.

Falsification tests

We performed falsification tests of Law and Marks's findings, in addition to the robustness test performed by Law and Marks (Tables 4 and 5 in Law and Marks 2009, 361-362). The goal of our falsification tests is to explore the possibility that

11. Registered nurses only in California and Massachusetts were represented by unions in the beginning of the 20th century. The California Nurses Association and the Massachusetts Nurses Association were formed in 1903.

Law and Marks's positive findings are simply spurious correlations. The approach we use is common in empirical analysis and includes several steps. First of all, we replicated their findings for engineers, registered nurses, practical nurses, pharmacists, and plumbers, for whom Law and Marks found positive and statistically significant effects. Our results differed from Law and Marks's findings for black practical nurses. The effect was negative and statistically insignificant instead of being positive and significant. Our other results were similar to those obtained by Law and Marks.

In the next step, we estimated a series of regressions where for every profession we used non-matching regulations. The idea was to determine whether we could find positive and statistically significant effects of licensing regulations if we used regulations that were unrelated to the profession of interest. Positive and significant findings would mean that the data was too noisy. First we estimated how licensing regulations that were introduced in engineering, nursing, and pharmaceutical professions affected minorities in the plumbing profession and vice versa giving us a total of 16 regressions. We found that non-matching regulations from other professions had positive and statistically significant effect on females in 38 percent of the cases (our Table 4, Panel A.). This number should be at most 10 percent.

TABLE 4. Results of the simulation analysis

Panel A.

Profession	Examined under engineers' regulations	Examined under registered nurses' regulations	Examined under practical nurses' regulations	Examined under pharmacists' regulations	Examined under plumbers' regulations
Engineers	xx	x	x	x	0.12
Registered Nurses	0.40	xx	x	x	x
Pharmacists	0.15	0.25	x	xx	x
Plumbers	0.12	x	0.34	x	xx

Note: Correlation of the non-matching and the original regulations are provided for cases with positive and statistically significant findings. x - indicates cases with non-significant findings. xx - indicates cases where the regulations match the professions.

Panel B.

Profession	Number of simulations	Percentage of cases with positive and statistically significant findings	Correlation of randomly generated and real regulations for cases with positive and significant findings		
			Average	Min.	Max.
Engineers	50	40%	0.27	0.06	0.52
Registered Nurses	50	12%	0.21	0.17	0.29
Pharmacists	50	26%	0.20	0.05	0.33
Plumbers	50	36%	0.21	−0.05	0.43

Additionally, we estimated a series of fifty regressions for each profession with randomly generated regulations. These regulations had positive and statistically significant effects on females in more than 10 percent of the cases in each profession and varied from 12 to 40 percent (see Table 4, Panel B.). Based on these findings we conclude that, consistent with our other criticisms, the data used by Law and Marks in the analysis was of low quality, and the positive effects that they found were most likely spurious correlations.

The results of our falsification tests can be replicated by using the Stata code provided online **here.**

Alternative evidence ignored by Law and Marks

There is much textual evidence that some of the leaders and interest groups behind licensing had discriminatory intent. Law and Marks (2009, 352, 353) acknowledge that some scholars have made this claim and briefly explain the reasoning behind it, but they do not actually acknowledge that this sort of textual evidence may be mounted. After summarizing their statistical results they say: "Hence, the conventional wisdom about how licensing affects minorities is not well supported, at least during the Progressive Era" (364). The implication is that, because the conventional wisdom is not supported by their statistical analysis, it is not well supported. No other evidence is considered.

Plumbing is one of the occupations included in their study. Law and Marks never consider the abundance of evidence like the following 1905 letter in *Plumbers' Journal* about a state licensing law that would eliminate blacks from the occupation:

> The Negro is a factor in this section, and I believe the enclosed Virginia state plumbing law will entirely eliminate him and the impostor from following our craft and I would suggest to the different locals that if they would devote a little time and money they would be able to secure just as good if not a better law in their own state. (quoted in Bernstein 1994, 96-97)

Many other samples of such discriminatory intent are easily found (see Bernstein 2001, ch. 2; Williams 2011, ch. 5; and Freeman 1980, 166-167). As for discrimination against women, in Arkansas and Georgia women were denied access to the bar until 1911 (Abel 1989, 69).

Law and Marks (353-354) cite several studies that indicate adverse effects on minorities (including Frech 1975; Sorkin 1977; Dorsey 1980 and 1983; and

Federman et al. 2006). But Law and Marks discount those studies, saying that their own study "allows us to speak more generally than previous studies about the impact of licensing on minority representation" (354). When Law and Marks come to their scholarly judgment, any such textual evidence of discriminatory intent, as well as the other evidence contained by the cited studies, seems to count for very little.

Another curious omission from the Law and Marks article is Freeman's chapter in the book *Occupational Licensure and Regulation* (Rottenberg, ed., 1980), a landmark in scholarship on licensing. In that chapter, "The Effect of Occupational Licensure on Black Occupational Attainment," Freeman conducts a statistical study of "the U.S. South in the period from the 1890s to 1960, during which white-dominated southern states often enacted or applied licensure laws discriminatorily" (165). Freeman says that his findings suggest "that, during the period of black disenfranchisement in the South, occupational licensure was a reasonably successful tool for reducing black employment in craft jobs, roughly as intended by many of its proponents" (170).

Law and Marks ignore the theoretical debate over licensing

Law and Marks discount the critical literature in another way. They write: "Finally, unlike the existing literature, our study has a clearly articulated alternative hypothesis. Theoretically, licensing regulation may increase the presence of minorities in occupations for which information about worker quality is an issue" (354).

Law and Marks act as though the quality-assurance rationale is something that the critical literature has neglected. In their introduction, after briefly treating the critical literature, they write:

> However, this [rent-seeking, etc.] is not the only role that licensing may play. Since Arrow (1963), economists have recognized that licensing can help solve informational asymmetries about occupational quality (Akerlof 1970; Leland 1979; Law and Kim 2005). If uncertainty about worker quality gives rise to statistical discrimination over observable characteristics like sex or race, then licensing regulation that serves as an imprimatur of quality can increase the presence of minority workers in regulated occupations…. (Law and Marks 2009, 352)

Thus, Law and Marks invoke the official rationale for licensing. The right way to view the critical literature is as a revisionist interpretation of licensing. The critical

literature builds on and pivots off the theoretical perspective highlighted by Law and Marks.

The critical literature is powerful because, while it accepts that trust and assurance are real problems, it so often shows that occupational licensing is not necessary to ensure trustworthiness—there are many private, voluntary practices and institutions for doing so—and that licensing requirements are often ill-suited to ensuring ongoing trustworthiness. Milton Friedman (1962, ch. 9) prominently articulated the case that licensing, while reducing freedom and imposing significant costs, achieves little in the way of quality and safety assurance that cannot be achieved by voluntary state certification. In discussion of reform options, Friedman focuses on the step from licensing to voluntary state certification. Under the certification system people are free to practice without a state certificate. Licensure, in contrast, bans practice until government permission is conferred on the individual.

Although one would never know it reading Law and Marks, the idea that voluntary seals of approval can work well, while avoiding many of the bad consequences of abridging the freedom of contract, has been noted by authors they cite. Hayne Leland concedes: "Under certification buyers have a wider range of choices…they can buy low-quality goods or services if they wish" (Leland 1980, 283). In his famous "Lemons" paper, George Akerlof (1970) discusses examples of "Counteracting Institutions" that undo the model; one such institution mentioned is licensing but he also discusses voluntary methods including guarantees, brand names, and chain stores. Finally, Kenneth Arrow (1963), too, repeatedly acknowledges shortcomings of licensing, including shortcomings in relation to certification (five pertinent quotations are gathered in Svorny 2004, 291).

Law and Marks also give no space to the idea that the demand for assurance elicits a supply of assurance (Klein 2012, ch. 12), and they show little concern as to whether licensing requirements are actually effective in ensuring ongoing trustworthiness. They essentially ignore the theoretical debate over licensing, a debate that goes back at least to Adam Smith's vehement opposition to such privileges (e.g., Smith 1776, 138). Instead, Law and Marks merely affirm the official rationale for licensing.

Furthermore, in many cases of licensure, it is unclear that Law and Marks's signaling theory fits actual practice. First, some economists argue that success in fulfilling all licensing requirements is unrelated to the skills needed for successful job performance (Dorsey 1983, 179; Gellhorn 1976, 12-18; Barton 2001 and 2003; Summers 2007). Second, the empirical evidence on licensure is mixed over whether licensure increases or decreases the quality received by consumers (Kleiner 2006; Pagliero 2008; Barker 2008; Powell and Vorotnikov 2011). Also, it is clear that for some requirements licensing is redundant as a quality signal. Law and Marks find

the introduction of the requirement of a high school diploma for teachers helped blacks. Presumably the diploma itself would serve as the signal of quality. The same applies to requiring a four-year medical degree for a physician's license.

Conclusion

The article by Law and Marks suggests that licensure did not tend to affect minorities adversely, even that it may have tended to benefit them. Their statistical investigation is rife with problems. First, the data upon which it is based includes both licensed and unlicensed practitioners, but we have strong evidence that in licensing states unlicensed practice was widespread. A fallback assumption that Law and Marks would perhaps invoke, that whites and blacks had similar rates of unlicensed practice (as well as men and women), is untenable because blacks surely had much higher rates of unlicensed practice. Hard numbers from a 1953 Maryland Senate investigation about plumbers show that only 2.1% of black Census-reported plumbers were licensed, whereas 51.8% of white plumbers were licensed. The example might be an extreme case, but it exemplifies the reality that rates of licensing among practitioners were usually higher, and probably much higher, for whites than for blacks, and probably likewise for men than women. This matter alone—which Law and Marks never acknowledge—could explain why their method would produce a spurious result that licensing did not typically harm minorities and sometimes even helped them.

We have made numerous additional criticisms: There is doubt about the accuracy of their data on when occupations were licensed. The criteria for including an occupation in the study suffer from a sample selection bias. We have catalogued at least eight problems with the analysis of their data, and several of the problems would create a bias in the direction of finding occupational licensing not detrimental to minorities. The problems could well have led Law and Marks to a result of "no finding" in cases in which the actual impact on blacks and women were negative.

As for the cases where Law and Marks found a positive impact, Table 5 provides a checklist as to whether the analytical problem gives reason for doubt. Law and Marks claim that we should have confidence in their findings, despite very low minority participation rates in some occupations, since the findings are similar across a range of occupations that adopted licensing at different times. In only three cases, however, did they have a positive finding that did not come from a small sample size. In two of these cases, those of black practical nurses and female registered nurses, we (and Law and Marks) know that their "licensing" data was mere certification, not licensure. The other case, that of black teachers,

also involves a conflation of certification and licensing as well as the concern that government was the main employer. We thus find it hard to argue that Law and Marks had a single meaningful positive finding, let alone a reliably consistent finding that would allow them to speak generally about the impact of occupational licensure on minorities.

We have also applied falsification tests to their findings, finding that randomly generated dates of regulation were able to achieve positive and significant findings for minorities in enough cases to indicate that Law and Marks's positive results are likely the result of similar spurious correlations. Unless better evidence is offered, there is no reason to abandon the conventional view that licensure generally harms minorities.

TABLE 5. Problems for the "positives" in Law and Marks

	Positive finding						
	Blacks			Women			
Problem	Prac. Nurses	Physicians	Teachers	Engineers	Pharmacists	Plumbers	Reg. Nurses
Enforcement is not perfect	X	X	X	X	X	X	X
Small participation rate		X		X	X	X	
Certification is not licensing	X		X	X			X
Licensing requirements varied/light	X			X	X	X	X
This "minority" dominated							X
Government employed most			X				
Unions may be significant			X			X	X

References

Abel, Richard L. 1989. *American Lawyers*. New York: Oxford University Press.

Akerlof, George A. 1970. The Market for "Lemons": Quality Uncertainty and the Market Mechanism. *Quarterly Journal of Economics* 84: 488-500.

Alchian, Armen A., and Reuben A. Kessel. 1962. Competition, Monopoly, and the Pursuit of Pecuniary Gain. In *Aspects of Labor Economics*, ed. H. Gregg Lewis, 157-184. Princeton: Princeton University Press.

Arrow, Kenneth J. 1963. Uncertainty and the Welfare Economics of Medical Care. *American Economic Review* 53: 941-973.

Barker, David. 2008. Ethics and Lobbying: The Case of Real Estate Brokerage. *Journal of Business Ethics* 80(1): 23-35.

Barton, Benjamin H. 2001. Why Do We Regulate Lawyers?: An Economic Analysis of the Justifications for Entry and Conduct Regulation. *Arizona State Law Journal* 33: 429-490.

Barton, Benjamin H. 2003. An Institutional Analysis of Lawyer Regulation: Who Should Control Lawyer Regulation—Courts, Legislatures, or the Market? *Georgia Law Review* 37: 1167-1250.

Beito, David T. 2000. *From Mutual Aid to the Welfare State: Fraternal Societies and Social Services, 1890-1967*. Chapel Hill: University of North Carolina Press.

Bernstein, David E. 1994. Licensing Laws: A Historical Example of the Use Of Government Regulatory Power Against African-Americans. *San Diego Law Review* 31(Feb.): 89-104.

Bernstein, David E. 2001. *Only One Place of Redress: African Americans, Labor Regulations, and the Courts*. Durham, N.C.: Duke University Press.

Council of State Governments. 1952. *Occupational Licensing Legislation in the States*. Chicago: Council of State Governments.

Dorsey, Stuart. 1980. The Occupational Licensing Queue. *Journal of Human Resources* 15: 424-434.

Dorsey, Stuart. 1983. Occupational Licensing and Minorities. *Law and Human Behavior* 7: 171-181.

Federman, Maya N., David E. Harrington, and Kathy J. Krynski. 2006. The Impact of State Licensing Regulations on Low-Skilled Immigrants: The Case of Vietnamese Manicurists. *AEA Papers and Proceedings* 96(2): 237-241.

Frech, H. E. 1975. Occupational Licensing and Health Care Productivity: The Issues and the Literature. In *Health Manpower and Productivity*, ed. John A. Rafferty, 119-142. Lexington, Mass.: D.C. Heath & Co.

Freeman, Richard B. 1980. The Effect of Occupational Licensure on Black Occupational Attainment. In *Occupational Licensure and Regulation*, ed. Simon Rottenberg, 165-179. Washington, D.C.: American Enterprise Institute.

Friedman, Milton. 1962. *Capitalism and Freedom*. Chicago: University of Chicago Press.

Gellhorn, Walter. 1976. The Abuse of Occupational Licensing. *University of Chicago Law Review* 44(6): 6-27.

Goldhaber, Dan D., and Dominic J. Brewer. 2000. Does Teacher Certification Matter? High School Teacher Certification Status and Student Achievement. *Educational Evaluation and Policy Analysis* 22(2): 129-145.

Hanushek, Eric. 2011. The Economic Value of Higher Teacher Quality. *Economics of Education Review* 30(3): 466-479.

Higgs, Robert. 1980. *Competition and Coercion: Blacks in the American Economy, 1865-1914*. Chicago: University of Chicago Press.

IPUMS-USA. 2010. Steven Ruggles, J. Trent Alexander, Katie Genadek, Ronald Goeken, Matthew B. Schroeder, and Matthew Sobek. *Integrated Public Use Microdata Series: Version 5.0* [Machine-readable database]. Minneapolis: University of Minnesota.

Kessel, Reuben A. 1958. Price Discrimination in Medicine. *Journal of Law and Economics* 1: 20-53.

Kessel, Reuben A. 1970. The AMA and the Supply of Physicians. *Law and Contemporary Problems* 35: 267-283.

Klein, Daniel B. 2012. *Knowledge and Coordination: A Liberal Interpretation*. New York: Oxford University Press.

Kleiner, Morris M. 2006. *Licensing Occupations: Ensuring Quality or Restricting Competition?* Kalamazoo, Mich.: W. E. Upjohn Institute for Employment Research.

Law, Marc T., and Sukkoo Kim. 2005. Specialization and Regulation: The Rise of Professionals and the Emergence of Occupational Licensing Regulation. *Journal of Economic History* 65(3): 723-756.

Law, Marc T., and Mindy S. Marks. 2009. Effects of Occupational Licensing Laws on Minorities: Evidence from the Progressive Era. *Journal of Law and Economics* 52: 351-366.

Law, Marc T., and Mindy S. Marks. 2012. Certification vs. Licensure: Evidence from Registered and Practical Nurses in the United States, 1950-1970. Working paper. **Link**

Leland, Hayne. 1979. Quacks, Lemons and Licensing: A Theory of Minimum Quality Standards. *Journal of Political Economy* 87: 1328-1346.

Leland, Hayne. 1980. Minimum-Quality Standards and Licensing in Markets with Asymmetric Information. In *Occupational Licensure and Regulation*, ed. Simon Rottenberg, 264-284. Washington, D.C.: American Enterprise Institute.

Lieberman, Myron. 2007. *The Educational Morass*. Lanham, Md.: Rowman and Littlefield Education.

Marshall, Ray. 1965. *The Negro and Organized Labor*. New York, London, and Sydney: John Wiley & Sons, Inc.

Northrup, Herbert R. 1944. *Organized Labor and the Negro*. New York and London: Harper & Brothers Publishers.

Pagliero, Mario. 2005. *What Is the Objective of Professional Licensing? Evidence from the US Market for Lawyers*. Ph.D. thesis, Department of Economics, London Business School.

Powell, Benjamin, and Evgeny Vorotnikov. 2011. Real Estate Continuing Education in Massachusetts: Rent Seeking or Consumer Protection? *Eastern Economic Journal* 38: 57-73.

Rottenberg, Simon, ed. 1980. *Occupational Licensure and Regulation*. Washington, D.C.: American Enterprise Institute.

Smith, Adam. 1976 [1776]. *The Wealth of Nations*, 2 vols. New York: Oxford University Press.

Snyder, Thomas D., and Sally A. Dillow. 2011. *Digest of Education Statistics: 2010*. Washington, D.C.: National Center for Education Statistics.

Sorkin, Alan L. 1977. *Health Manpower: An Economic Perspective*. Lexington, Mass.: Lexington Books.

Spero, Sterling D., and Abram L. Harris. 1931. *The Black Worker: The Negro and the Labor Movement*. New York: Columbia University Press.

Stephenson, E. Frank, and Erin E. Wendt. 2009. Occupational Licensing: Scant Treatment in Labor Texts. *Econ Journal Watch* 6(2): 181-194. **Link**

Summers, Adam B. 2007. Occupational Licensing: Ranking the States and Exploring Alternatives. *Policy Study* 361. August. Reason Foundation (Los Angeles). **Link**

Suzuki, Peter. 1985. Vernacular Cabs: Jitneys and Gypsies in Five Cities. *Transportation Research A* 19: 337-347.

Suzuki, Peter. 1995. Unregulated Taxicabs. *Transportation Quarterly* 49(Winter): 12-38.

Svorny, Shirley. 2004. Licensing Doctors: Do Economists Agree? *Econ Journal Watch* 2(1): 279-305. **Link**

Troy, Leo. 1965. *Trade Union Membership, 1897-1962*. New York: National Bureau of Economic Research.

Vorotnikov, Evgeny S. 2011. The Adverse Consequences of Entry Regulation in the Legal Profession. Working paper.

Wilgus, Horace L. 1908. Legal Education in the United States. *Michigan Law Review* 6(8): 647-682.

Williams, Walter E. 1982. *The State Against Blacks*. New York: McGraw Hill.

Williams, Walter E. 2011. *Race and Economics: How Much Can Be Blamed on Discrimination?* Stanford, Calif.: Hoover Institution Press.

About the Authors

Daniel Klein is professor of economics at George Mason University and chief editor of *Econ Journal Watch*. His email address is dklein@gmu.edu.

Benjamin Powell is associate professor of economics at Suffolk University and a senior fellow at the Independent Institute. His email address is benjaminwpowell@gmail.com.

Evgeny Vorotnikov is a post-doctoral fellow at the University of Minnesota in the Department of Applied Economics and a visiting scholar at the Carlson School of Management and the Minnesota Population Center. His email address is evvmail@gmail.com.

Marc Law and Mindy Marks' reply to this article
Go to Archive of Comments section
Go to September 2012 issue

EJW
ECON JOURNAL WATCH
Scholarly Comments on Academic Economics

ECON JOURNAL WATCH 9(3)
September 2012: 234-255

Occupational Licensing and Minorities: A Reply to Klein, Powell, and Vorotnikov

Marc T. Law[1] and Mindy S. Marks[2]

LINK TO ABSTRACT

Introduction

In May 2009, the *Journal of Law and Economics* (*JLE*) published our article titled "Effects of Occupational Licensing Laws on Minorities: Evidence from the Progressive Era." In that article, we investigated the impact of the adoption of state-level occupational licensing regulation on the participation of minority workers in a range of skilled and semi-skilled occupations. Specifically, we took advantage of quasi-experimental variation, afforded by the fact that different states adopted occupational licensing regulation at different times, to identify the effect these laws had on the prevalence of female and black workers in eleven different occupations using a differences-in-differences framework. We found that the adoption of these laws did not reduce minority participation in most occupations. In fact, for many occupations, we found that the adoption of licensing laws was correlated with increases in minority participation. We argued in our article that the evidence presented was generally inconsistent with the received wisdom, which claims that licensing laws reduced minority participation in most occupations. Instead, the evidence is more supportive of an alternative hypothesis that posits

1. University of Vermont, Burlington, VT 05405.
2. University of California, Riverside, Riverside, CA 92521.

that licensing may help minorities, particularly in occupations for which information about worker quality is difficult to determine.

In "Was Occupational Licensing Good for Minorities? A Critique of Marc Law and Mindy Marks," Daniel Klein, Benjamin Powell, and Evgeny Vorotnikov (henceforth KPV) take issue with our findings. They argue that our study is "rife with problems" (KPV 2012, 228). Among other things, they believe that: (i) the data we use (individual-level census returns) are inappropriate for addressing the question because of imperfections in how licensing laws were enforced; (ii) there is sample selection bias in the set of occupations we have chosen to analyze; (iii) several of our findings are based on small numbers of minority workers; (iv) we have conflated licensing and certification in some instances; (v) there is measurement error in how we measure the timing and restrictiveness of occupational licensing; and (vi) the omission of controls for interstate migration and craft unionization bias our results in favor of our findings. KPV therefore conclude: "[T]here is no reason to abandon the conventional view that licensure generally harms minorities" (2012, 229).

We welcome this opportunity to join KPV in a discussion of our study. However, after reviewing their arguments, we remain convinced that our study properly identifies the effect of licensing on minority participation during the period under investigation. While we concede that the interpretation of our evidence may be slightly altered for some occupations (where, we agree, that we have lumped together licensing and certification), and that the nature of the data make it difficult to identify the precise channel(s) through which licensing impacts minority participation, KPV's criticisms do not definitively show that our methodology is biased in favor of our findings. In fact, nowhere in their critique do KPV show that any of our original regression findings are altered by the inclusion of new variables or changing variable definitions. Indeed, many of KPV's concerns about measurement error bias our empirical strategy *against* finding a positive effect of licensing on minority participation, thereby strengthening our original claims.

Before proceeding with our rebuttal, we feel we should make a few disclosures of a more personal nature. First, as empirical economists, we did not have strong priors regarding the impact that licensing laws should have on minority participation in various occupations. Different hypotheses have different predictions regarding the direction of impact. An important task of empirical scholarship in economics is to determine the effect that policy has on economic outcomes as objectively as possible. This was our only goal in conducting this study. Our preference is simply to allow the data to speak and to base our conclusions on the data. Second, while we are interested in knowing the direction of effect, we have no stake in the outcome of this debate with respect to current policy. We therefore take issue with KPV's claim that we are in favor of licensing

today (see KPV 2012, 222 n. 10, which claims that we give policy advice today based on our findings). Nowhere in our paper (nor in our other writings) do we ever argue that licensing regimes should be extended. In our conclusion (Law and Marks 2009, 364) we did suggest that licensing may help minority workers signal quality in some cases, but this hardly amounts to an argument in favor of current licensing regimes. We therefore ask KPV not to make inferences regarding our views about policy from our paper. Third, the fact that licensing may not have been very harmful on net for minority participation during the Progressive Era does not mean that it has not been harmful at other times and in other places or along other dimensions or that some minority workers were not harmed.[3] Our study only addresses the impact of licensing within certain occupations over a given period. We can say nothing out-of-sample, and are well aware of other studies that have found harmful effects on minorities in other occupations and time periods (see, for instance, Federman, Harrington, and Krynski 2006).

Finally, the reader may be interested to know something of the background to this article. In March 2012 KPV submitted a critique of our 2009 article to the *JLE*. By request of the editor of the *JLE*, KPV also sent us a copy of their critique (cited hereafter as Vorotnikov et al. 2012). The following month we responded to KPV's critique and also sent a copy to the *JLE* (this reply is cited hereafter as Law and Marks 2012). The *JLE* subsequently rejected KPV's critique (which, we understand, was a revised version of the critique they initially sent to us). Since then, KPV have sent their critique to *Econ Journal Watch* (*EJW*). The present article is a revised version of our original response to KPV. We note that KPV have amended their critique to incorporate some of our replies. Accordingly, in addition to replying to each of the criticisms leveled in the *EJW* version of their critique, we will include a discussion of some of KPV's original criticisms.

KPV have organized their critique of our paper in four sections. The first section details problems that KPV have with our data. The second section outlines problems that KPV have with our analysis of the data. The third concerns alternative qualitative and historical evidence KPV believe we have ignored. The last section deals with the theoretical debate over licensing. We organize our response along the same lines.

3. Indeed, in an earlier article on the emergence of Progressive Era occupational regulation, one of us (along with Sukkoo Kim) wrote: "We hope that future scholars will take up the task to determine if and when licensing regulations became a tool to advance the narrow interests of professionals at the expense of the general public" (Law and Kim 2005, 754).

KPV's problems with the data

KPV make two claims about the why the data we use are inappropriate for investigating the effect of licensing on minority participation. The first is that census-reported practitioners in a licensed state are not necessarily licensed. Imperfections in how licensing laws are enforced may therefore bias our results because our data possibly include female and black practitioners who self-declare to be within an occupation, even though they are practicing without a license. KPV therefore, in the earlier version of their critique, accuse us of "assuming that licensing restrictions were perfectly enforced" (Vorotnikov et al. 2012, 3) and argue that our preferred interpretation of our results requires this assumption. Relatedly, they also present qualitative historical evidence suggesting that enforcement of licensing standards was weak during the period under investigation (especially for blacks) and that unlicensed practitioners often practiced in racially segregated markets. KPV's second claim is that the occupational licensing data we use are flawed. Specifically, they are concerned that the dates of initial licensing laws are measured with error. As evidence of this, KPV examine data on the occupational licensing laws for lawyers from other sources and show that the our primary data source on the timing of licensing laws contains errors with respect to the dates of initial licensing laws for the legal profession.

Is imperfect enforcement a source of bias?

Let us start with the first issue. It is true that the census does not identify whether or not an individual has a license. To our knowledge, there are no systematic data on the licensing status of individuals across a broad range of occupations during this period. Occupational status from individual census returns is self-declared. Accordingly, a person who declares herself a plumber will be coded as a plumber, regardless of whether or not the individual possesses a license to practice plumbing. We do not dispute this fact. The question for us is whether this makes it impossible to correctly identify the effect of licensing on minority participation in various occupations using census data. In other words, do imperfections in the enforcement of licensing laws render our empirical strategy invalid?

We believe that they do not. For one thing, if imperfect enforcement is to systematically bias our results, it is probably classical measurement error that biases our estimates toward zero (i.e., attenuation bias). KPV present no *systematic* evidence to suggest that imperfections in enforcement would be a source of non-

classical measurement error. The evidence they present—on plumbers in Maryland in the 1950s—is hardly fatal since it represents only a single occupation in a single year whereas our analysis covers eleven occupations in 48 states over seven census years.

Second, while KPV claim that we do not mention the fact that a practitioner in a licensed state is not necessarily licensed, the empirical strategy we used already acknowledges the possibility of imperfections in regulatory enforcement due to grandfathering. As a robustness check, we re-estimated all of our key regressions restricting the sample to young workers (see Tables 4 and 5 in Law and Marks 2009, 361-362). Licensing laws, when applied, are more likely to be binding (and hence, enforced) for young workers (who are new to a profession) than to older workers (who are incumbents and already have established themselves). The fact that we find similar results for young black and young female workers as for the whole sample suggests that imperfections in enforcement are not an important source of bias.

Finally, it is standard practice in the literature on the labor market effects of state-level public policy to use micro data from government sources like as the U.S. Census or the Current Population Survey. In all of these data sources, important variables like occupation and earnings are self-reported. Even when it is known that the state-level public policy is not perfectly enforced, researchers estimate models using a similar research design as ours. For instance, state-level minimum wages are not perfectly enforced. Enforcement may be uneven across different demographic groups, and yet economic analysis of the effects of minimum wages on labor market outcomes use empirical specifications that are, in spirit, much like ours.[4] If we were to take KPVs criticism on this score seriously, we would also have to question the validity of this entire literature.

Accordingly, we maintain that our data set and empirical strategy allow us to correctly identify the effect of licensing on minority participation. Indeed, since our key empirical question is how state-level licensing affects minority labor force participation in a variety of occupations, the empirical model is the correct one, even if it is harder or impossible for minorities to gain a license. We agree, however, that the approach may not allow us to determine the precise causal mechanism. This is a common problem for quasi-experimental approaches to causal inference. Minority participation in some occupations may have increased as a result of licensing because of the signaling value of having a license, which is our interpretation of the evidence. On the other hand, minority participation may increase because licensing raises prices and creates a black market in which minority

4. See, for instance, Neumark and Wascher (2001) and Acemoglu and Pischke (2003).

workers are disproportionately represented. Our research design does not allow us to cleanly distinguish between these two alternatives.

That said, there are good reasons to be skeptical of the second interpretation. If licensing creates an illicit market that benefits minorities, why is it that we only observe a positive effect of licensing on minority participation in some occupations but not for others? *A priori*, it is not obvious why licensing should facilitate an illicit market for female engineers, plumbers, and pharmacists as well as for black teachers and doctors, but not for black barbers or beauticians. In fact, our results show that black participation in the barbering profession was reduced by occupational regulation of barbers, which raises the question of why black barbers were unable to operate underground but black teachers and doctors were? Our interpretation, we believe, is more consistent with the pattern of positive and negative results (i.e. licensing increased minority participation in professions where practitioner quality was harder to ascertain and where statistical discrimination more likely, and licensing reduced participation in occupations where quality was not hard to ascertain and where minority workers were a viable competitive threat), but we agree that the evidence in favor of this interpretation is not definitive.

Finally, KPV argue that our interpretation hinges on perfect enforcement. We disagree with this claim. In order for licensing to reduce statistical discrimination it must be the case that *some* potential customers value the information provided by a license and that *some* minorities be able to obtain one. No assumption of perfect enforcement is required.

Are the data on when occupational licensing laws were adopted flawed?

KPV's evidence for this claim is based on their analysis of a *single* occupation. They present no evidence showing that the data on when other occupations became regulated are systematically misreported. In fact, we excluded lawyers—the one profession for which they show that the data on licensing are in error—precisely because the quality of the data was weak (information on the timing of licensing was reported as unknown for 19 states in the Council of State Governments' 1952 study). This significantly lessens the force of the critique.

Nonetheless, we recognize that data on the timing of licensing laws for the occupations that we do examine may be recorded with error in the Council of State Governments' publication. Once again, the key question for us is whether this source of measurement error systematically biases our findings. In order for measurement error in the timing of licensing laws to bias our results, it would have to be correlated with trends in minority representation within an occupation over time. There are no reasons to believe that this is the case, nor do KPV present

any data to show this. Hence, we remain skeptical that this is an important source of systematic bias. Indeed, random measurement error in the reported timing of licensing, which is likely to be the relevant kind of measurement error in this context, will cause attenuation bias, making it *harder* to find that licensing has a statistically significant effect.

Nevertheless, our study has two features that help us deal with possible measurement error over the timing of licensing laws. First, we excluded from analysis any state for which the year in which licensing was adopted was unknown. Second, because we use census data (which are reported every decade), our empirical strategy allows for mis-measurement within a census decade.

KPV's problems with our analysis

In Table 5 (KPV 2012, 229), KPV list six additional problems with our empirical analysis.[5] The first concerns which occupations were included for analysis. The second is that we have in some cases conflated licensure with certification. The third is that we do not include measures of how restrictive licensing regimes were. The fourth is that we include one occupation (teachers) where most employment was by the public sector. The fifth is that it is hard to interpret the results for occupations like nursing, where most workers were women. The sixth is that there is omitted variable bias because we do not control for the presence of craft unions or trends in interstate migration.

Occupations covered in our study

KPV have three complaints about the sample of occupations that we analyze in our study. First, KPV believe that our requirement that at least one percent of an occupation had to be either female or black to be included in our analysis biases our overall findings because we excluded occupations for which there are very small numbers of minority workers. They argue that are the requirement that at least one percent be back or female "could well have selected certain occupations out of the study *because* licensing was so strongly discriminatory." (KPV 2012, 218). Second, in the earlier version of their critique, KPV argue that there are other occupations that we could have included but we did not. Using our criteria for selection, KPV argued that dentists, insurance brokers, and real estate agents should also have

5. Interestingly, KPV only have problems with the occupations for which we find that licensing had a positive effect on minority participation. They seem untroubled by our analysis of barbers, where we find a negative effect using the same empirical methodology.

been included (Vorotnikov et al. 2012, 9). The implicit suggestion is that we have cherry-picked those occupations we included in order to bias our results in favor of finding a positive impact of licensing on minority participation. For reasons that are undisclosed, KPV have chosen to omit this particular criticism in their updated critique, perhaps because we dealt with it successfully in our original reply (Law and Marks 2012, 10). Third, KPV argue that because minority participation rates in many of the occupations that we studied were very low and close to the one-percent threshold, our results are spurious (KPV 2012, 219).

There is an inherent contradiction in KPV's first and third criticisms. On the one hand they claim that our one-percent requirement results in too many occupations being excluded. On the other hand, they say that our one-percent requirement results in too many occupations being included. KPV seem to want to have their cake and eat it too. While we agree that it is hard to know what the appropriate participation cutoff should be, and we acknowledge there is a tradeoff, the fact of the matter is that some cutoff must be chosen. For many of the occupations for which we have data, minority participation was very low during this period, rendering statistical analysis problematic. KPV provide no remedy, nor do they definitively show that our choice of a one-percent cutoff systematically biases our results in favor of finding that licensing increases minority participation.

Now to KPV's second concern, which, as noted, is not mentioned in the current version of their critique. First, let us state most emphatically that we *did not* cherry-pick which occupations to examine. Our only goal was to find a sample of occupations that represented a broad range of skills and for which there was enough minority participation to conduct an empirical analysis. As we write in the introduction to this reply, we had no strong priors about whether minority participation should be harmed or helped by occupational regulation, nor do we have any stake in the outcome of this debate. While eleven occupations do not cover all licensed occupations, they still constitute a non-trivial share of the non-agricultural share of the labor force during this time (six percent—see Law and Marks 2009, 352). Indeed, there are very few studies of the effects of licensing on minorities that examine more than one or two occupations at a time. Accordingly, our coverage is broader than most.

Let us now deal with the three occupations that KPV claimed that we overlooked. We excluded dentists because over the six census years for which we have data on dental licensing laws (1880 through 1940, excluding 1890, for which there are no IPUMS data), there were only 32 black and 37 female dentists (out of a total of 2,055 dentists). For what it is worth, when we do estimate the effect of licensing on black and female participation in dentistry using identical regressions as for the other occupations, the effects of licensing are not statistically significant.

Insurance brokers and real estate agents are two occupations that we could have included in the original paper but did not. Our failure to include these occupations was an oversight on our part, and we thank KPV for pointing this out. Accordingly, we estimated the effects of licensing of these two occupations on minority participation using the same empirical model that we used in our original study. Regression results are shown in Table 1. The coefficient of interest is the interaction between the minority and licensing indicator variables. For insurance agents, the adoption of licensing did not have a statistically significant effect on either female or black participation, while for real estate agents, licensing had a positive and statistically significant effect on female and black participation. Hence, if we had included these occupations in our original study we would have found more evidence in favor of the view that licensing did not generally harm minority participation and in some cases even helped.

TABLE 1. Effect of occupational licensing on minority workers, by occupation

	Insurance Agents: Black	Insurance Agents: Female	Real Estate Agents: Black	Real Estate Agents: Female
Licensing indicator	−.048 (.039)	−.047 (.039)	−.012 (.042)	−.018 (.041)
Black x Licensing	−.053 (.011)		.376** (.054)	
Black	−.520** (.078)		−.712** (.071)	
Female x Licensing		−.034 (.070)		.204** (.041)
Female		−.550** (.052)		−.473 (.039)
Sample size	963,836	963,836	1,874,558	1,874,558
State-years	294	294	240	240
Years included	1870-1940	1870-1940	1910-1950	1910-1950
Minority representation	240 (4.99%)	240 (4.99%)	108 (2.25%)	379 (7.88%)

Note. Each column contains a separate regression. State and year fixed effects and individual- and household-level controls (age, gender, race, literacy, urban residence, domestic, married, widowed, children, two families, three families, at school) are included when available. Robust standard errors, clustered at state level, are in parentheses. * and ** denote statistical significance at the 5- and 1-percent levels, respectively. IL, MI, NC, and UT were excluded from the insurance agent sample due to missing licensing data. Minority representation has percentage of the occupation that is minority (black or female) in parenthesis as well as number of workers. Information on the introduction of state licensing laws is from the Council of State Governments (1952).

Licensing vs. certification

KPV claim that there are four occupations we study for which we have conflated licensure and certification: registered and practical nurses, teachers, and engineers. We agree with KPV that for some occupations we have conflated certification with licensure. Because the relevant table from the Council of State

Governments study was titled "Occupations Licensed by State and Dates of Licensing Statutes," we assumed that these referred to the dates at which states adopted mandatory licensure. However, our overall conclusion—that licensing aids minority representation in some cases—is not altered by this fact for three of the four occupations.

First, consider teachers. KPV write that "in no states were people required to have a license to teach in a *private* school" (2012, 219, emphasis added). KPV furnish no evidence, however, to show that licensing was not in place for *public* school teachers. Later on, KPV (2012, 221) mention that in 1919-20, 93 percent of teachers worked in public schools. Accordingly, it is possible that in fact the vast majority of teachers were subject to licensure.

TABLE 2: Effect of occupational licensing on minority workers in nursing

	Registered Nurses: Black	Practical Nurses: Black	Registered Nurses: Female	Practical Nurses: Female
Licensing indicator	.006 (.021)	−.055 (.040)	.072 (.060)	−.153 (.095)
Black x Licensing	.281* (.103)	.012 (.081)		
Black	−.424** (.033)	.270** (.056)	−.291** (.076)	.274** (.044)
Female x Licensing			−.050 (.076)	.108 (.081)
Female	1.510** (.040)	1.109** (.056)	1.655** (.049)	1.077** (.069)
Sample size	1,196,523	1,165,341	1,196,523	1,165,341
State-years	96	82	96	82
Years included	1950-1960	1960-1970	1950-1960	1960-1970
"Minority" representation	505 (5.03%)	860 (19.78%)	10,036 (97.79%)	4,165 (95.81%)

Note. Each column contains a separate regression. State and year fixed effects and individual- and household-level controls (age, age squared, gender, race, education, size of metro area, domestic, married, widowed, number of children) are included. Sample is restricted to those in the labor force. Robust standard errors, clustered at state level, are in parentheses. * and ** denote statistical significance at the 5- and 1- percent levels, respectively. AR, CA and MT had licensing in place at the beginning of the sample and were excluded from the registered nursing analysis. AK, AR, CO, CT, FL, ID, LA, NV, NY and RI had licensing in place at the beginning of the sample and were excluded from the practical nursing analysis. "Minority" representation has percentage of the occupation that is minority (black or female) in parenthesis as well as number of workers. Information on the timing of licensure is from Monheit (1982).

Second, let us address the two nursing professions. We have more recently collected data on the timing of mandatory licensure for the two nursing professions. Specifically, we have data on registered nurses from 1950 to 1960 and practical nurses from 1960 to 1970. During this period, 21 states adopted mandatory licensing for registered nurses and 20 states adopted mandatory licensing for practical nurses. We can therefore estimate the effect of licensure on the participation on female and black registered and practical nurses during the period when licensing became mandatory. Regression results are displayed in

Table 2.[6] For women, mandatory licensing had no statistically significant impact on participation in either practical or registered nursing. For blacks, the adoption of mandatory licensing raised black participation in registered nursing but had no statistically significant effect on black participation in practical nursing. Hence, it would seem that when nursing licensure became mandatory, it still did not have significantly negative effects on participation, and likely increased the representation of blacks among registered nurses. Accordingly, the results from these regressions are consistent with our original findings.

Do we measure the restrictiveness of licensing?

KPV accuse us of including no measures of the restrictiveness of occupational licensing laws. Because the restrictiveness of licensing laws varied by state, KPV believe that failure to control for differences in the strictness of laws biases our results.

We have several replies to this argument. The first is that the claim that we do not have any measure of restrictiveness is patently false. We do. A whole section of our article (Section 6) studies the effects of specific licensing requirements on the prevalence of females and blacks among teachers and physicians. In fact, as we wrote in this section of our article, the motivation for collecting data on particular licensing requirements is precisely to reduce the problem of measurement error (Law and Marks 2009, 362).

KPV then argue that, for the two occupations that we study in greater depth, we have *selectively* chosen which requirements to include in our regression analysis. In discussing our analysis of physicians, KPV note that the two specific licensing requirements we included in our regressions—whether a four-year medical degree was required, or whether there were pre-medical education requirements—are both educational requirements. Since licensing requirements vary along many dimensions, they wonder why we chose to focus on these two requirements. The suggestion is that had we chosen other requirements, the results might be quite different.

Licensing requirements obviously do vary along many dimensions. However, educational requirements have been shown to be among the most prevalent (see Kleiner and Krueger 2009), so the focus on educational requirements would seem well justified, especially since teachers and doctors are among the more highly

6. Because the information provided by the decennial censuses changes over time, these regressions are not identical to those estimated in our original study. However, efforts were made to make them as similar as possible. For instance, in the analysis conducted using earlier census periods, we excluded military, retired, and housewives. The latter two are no longer reported categories. Hence, for this analysis, we excluded anyone who was not in the labor force.

skilled occupations in our sample. Additionally, if we are to use variation in licensing requirements to measure licensing, we need licensing requirements that varied sufficiently during the period under analysis. The other licensing requirements for which we could find data (for instance, whether or not a physician had to pass a licensing exam) did not vary significantly during this time. Hence, we could not use variation in these requirements to identify the effects of licensing.

For the remaining nine occupations in our sample, it is true that we do not have more direct measures of the restrictiveness of a state's licensing regime. The relevant question is therefore whether variation in restrictiveness is correlated with minority participation in such a way that states with low restrictiveness are those where minority participation is growing fastest. KPV provide no systematic evidence showing that variation in restrictiveness was correlated with minority participation. We have no reason to believe that the failure to control directly for restrictiveness induces anything other than classical measurement error, which should make it *harder* for us to find that licensing had any effect on minority participation.

Some occupations (teachers) are mostly employed by the public sector

KPV object to the inclusion of teachers in our analysis, largely because most teachers are public sector workers. We are not quite sure why this should matter. Our objective in conducting this study was merely to determine the effect of occupational regulation on minority participation. While we agree that the factors influencing employers' or customers' decisions may vary by sector, it is not clear to us that the net impact of licensing should be systematically different for the public sector. Licensing could potentially serve as either an entry barrier that facilitates discrimination against minorities or as signal of quality that reduces statistical discrimination, even in environments where the employer is the government and other (perhaps political) considerations are present. KPV have not clearly articulated why our overall findings are biased as a result of including public sector workers. Accordingly the inclusion of teachers among the occupations we analyze seems warranted.

Nurses and teachers?

KPV argue that it is not obvious how to interpret our findings regarding the effect of occupational licensing on teachers and nurses on women, two occupations that were (and still are) predominantly female. We agree and were up front

about this in our article (Law and Marks 2009, 363). We have nothing more to add on this issue.

Craft/trade unions and interstate migration as omitted variables?

KPV's final complaint in this section concerns omitted variable bias. The first omitted variable they are concerned about is the extent of craft or trade unionization in some occupations. During this period, some of the occupations we analyze (for instance, plumbers) were unionized in some states. In the current version of their critique, KPV argue that if craft unionization and licensing were *substitute* mechanisms for reducing minority participation in certain occupations, then failure to control for the presence of craft unions will bias our results toward finding that licensing helped minority participation.

While we agree that this would be the direction of bias if it were true that unionization and licensing were substitutes, we think it worth pointing out that this is precisely the opposite of what KPV wrote about craft unions in the earlier version of their critique (Vorotnikov et al. 2012, 15-17). In that earlier version, KPV noted that unions often used licensure as an additional tool for reducing minority participation in a given occupation.[7] Hence, the suggestion from the earlier version of their critique was that craft unions and licensure were *complementary* mechanisms for reducing minority and female participation. But in that case, if failing to control for unionization biases our estimates, it probably biases our estimates of the impact of licensing *against* our actual findings. Let us take KPV's original argument seriously. Suppose unionization is positively correlated with the adoption of licensure, perhaps because unions lobby for licensure as an additional entry barrier that facilitates racial discrimination. In our view, this seems plausible since unions are well positioned to solve the collective action problem that must be overcome to obtain licensing legislation in the first place. In that case, if we fail to control for unionization, we will *over-estimate* the *negative* impact that licensure has on minority participation. In other words, our regressions are biased *against* finding either no effect or a positive effect. We are unsure why KPV have chosen to reverse their original position on the relationship between licensure and craft unionization. One possibility is that they quickly discovered that their original

7. In fact, when describing the qualitative evidence about this relationship in their original critique, KPV wrote that unions "worked hand-in-glove with licensing requirements to keep blacks out." They also argued that unions "had a legacy of discrimination" and "often lobbied for licensure" (Vorotnikov et al. 2012, 16-17). This evidence is omitted in the current version of their critique. To be fair, in the earlier version of their critique, KPV also mentioned the possibility that unionization and licensure may have been substitute mechanisms, but the bulk of their discussion seemed to suggest the opposite.

position on this relationship was not helpful for their critique, which we pointed out in our original reply (Law and Marks 2012, 17). However, we can never uncover KPV's true reasons.

In fact, it remains mostly a conjecture as to whether unionization was even correlated with licensure during this period. Unfortunately, there are no reliable state-level data on unionization prior to the mid 1960s (see Hirsh, Macpherson, and Vroman 2001). However, some other evidence we presented in our article suggests that predictors of unionization are not correlated with the adoption of licensure. In our effort to show that the adoption of licensing constituted a valid quasi-experiment, we estimated a series of regressions to determine whether states that adopted licensing differed systematically from states that did not (Law and Marks 2009, 356). Among other things, we found that the growth rate of an occupation (a predictor of unionization) was uncorrelated with the adoption of licensing, which, indirectly, suggests that licensing was uncorrelated with unionization.

In the original version of their critique, KPV claimed that our results are biased because we failed to control for changes in the minority population (Vorotnikov et al. 2012, 17). This claim, as we pointed out in our original reply (Law and Marks 2012, 18) is false, because in the same series of regressions of the correlates of adoption, we also included measures of the minority population, finding no correlation between changes in the minority population (either within the occupation or within the labor force) and the adoption of licensure (Law and Marks 2009, 356). If changes in the minority population are uncorrelated with the adoption of licensure, not including this variable will not alter our results. KPV had therefore failed to identify an omitted variable that we need worry about and appeared oblivious to the efforts we had undertaken to show that the adoption of licensing during this period constitutes a valid quasi-experiment. In their updated critique, KPV now claim that we fail to control for interstate migration of minorities, but acknowledge that we included black and female shares of the labor force in our analysis of the adoption of licensing. We do not believe this to be a serious problem because black and female shares of the labor force subsume black and female interstate migrants. For this to be a source of bias, one would have to think that the effects of licensure on minority participation should be different for minorities born in state versus those born in other states. KPV present no evidence to suggest that this should be the case.

Falsification tests?

KPV conduct two different empirical exercises to explore if our positive findings are "simply spurious correlations" (KPV 2012, 224). The authors do not state what underlying factor(s) they suspect might be driving these spurious cor-

relations between state-level licensure and minority occupation representation. As noted earlier, we ruled out many potential factors including urbanization, and the share of the labor force that is minority when we tested for the validity of the quasi-experiment (see Law and Marks 2009, 356). Unfortunately, in their text, KPV provide little detail about the specifics of these falsification tests. Indeed, KPVs description is limited to two brief paragraphs. With only a week left before our submission deadline, KPV sent us their code. We have used this to make inferences about what exactly they did.

KPV's first falsification test involved estimating "how licensing regulations that were introduced in engineering, nursing, and pharmaceutical professions affected minorities in the plumbing profession and vice versa giving us a total of 16 regressions" (KPV 2012, 224). The results of this exercise are displayed in Table 4, Panel A of their critique (224). KPV argue that if a regression of the effects of, say, engineering regulation, on participation of women in the plumbing profession generates a positive and statistically significant effect, this shows that our methodology is biased. According to KPV, "false positives" of this sort are found in 6 of 16 cases.

We have several problems with KPV's first falsification test. First, it is unclear why KPV only estimated these regressions using females as the dependent variable. Second, we are uncertain why KPV have estimated 16 regressions, not 20. If there are five occupations (engineering, registered nursing, practical nursing, pharmacy, and plumbing), then participation in each occupation can be estimated using data on regulation for four other occupations. Practical nurses are omitted from the first column of Panel A. Why practical nursing is not included is unclear. Perhaps it was because KPV were not able to replicate our results for black practical nurses. In any event, KPV's reasons for omitting this occupation are left unexplained.

Third, the validity of KPV's first falsification test also requires that there be no correlation in the timing of regulation among these seemingly unrelated occupations. We strongly suspect that this is not the case. State-level factors such as population growth or the political party in the state house or of the state governor could result in multiple professions becoming regulated at the same time. If regulation of, say, engineers is correlated with the regulation of pharmacists, and if the regulation of pharmacists increases minority labor force participation in pharmacy, then regulation of engineers will serve as a weak proxy for the regulation of pharmacists.

Finally, it is unclear what numbers are presented in KPV's Table 4, Panel A. Take the 0.4 that is shown in the second row of the panel. One interpretation (which seems to be suggested by the note below Panel A) is that 0.4 is the correlation between registered nursing regulation and engineering regulation. If this

is the case, then our previous point is made for us: it would not be surprising if a regression of the effects of regulation of one profession has a statistically significant effect on participation in another occupation if the regulations are in fact correlated with each other. Hence, we reject the claim that "false positives" in these regressions prove that our methodology generates spurious results.

For their second falsification test, KPV "estimated a series of 50 regressions for each profession with *randomly* generated regulations" as the key independent variable (KPV 2012, 225, emphasis added). It appears that the analysis was conducted using female participation in each profession as the dependent variable. Once again we do not know why KPV did not do the same for blacks. In any event, according to KPV, a positive and statistically significant relationship between these randomly generated regulations and female participation is found more than 10 percent of the time for four professions (see KPV's Table 4, Panel B).

We have several problems with this falsification test. First, if the authors are truly using randomly generated regulations, then basic statistical reasoning suggests that with a sufficient number of runs, they should find an effect in no more than 10 percent of the cases. (Note, by the way, that they are using a 10 percent standard whereas we held ourselves to a five percent standard). One reason why KPV may find positive and statistically significant effects more than 10 percent of the time is that 50 runs is insufficient. It is possible that with a larger number of runs, the percentage of statistically significant cases would converge to 10 percent. In fact, KPV do not report boostrapped standard errors so we cannot rule out the possibility that the results are in fact significant only 10 percent of the time.

It is also possible that while KPV state that they *randomly* generated regulations, they did not really do so. From looking at their code, it appears that for each state they randomly assigned a treatment year. This strategy induces correlation across time that is not accounted for in their estimation strategy, which could result in spurious correlations being found in more than 10 percent of cases (Bertrand et al. 2004).

Finally, regardless of what KPV have done, the numbers in Table 4, Panel B (2012, 224) suggest that the average correlation between KPV's randomly generated regulations and the true regulations was between .20 and .27, which, in turn, implies that they are hardly random. Since KPV's randomly generated regulations appear to be correlated with the true regulations (which affect participation), it is not surprising that they find significant results in more than 10 percent of the cases.

In summary, insufficient detail is provided about these tests for KPV to claim that they serve as a "damaging" (KPV 2012, 211) critique of our empirical approach. Additionally, from what we can infer about what KPV have done (again, we emphasize that KPVs description of these tests is vague at best, and we only received their code one week prior to the deadline), we have multiple reasons to

be skeptical about whether these are in fact valid falsification tests. We contend that the implicit falsification tests we conducted on older workers who were grandfathered by occupational regulation and thus not affected by its adoption furnish a cleaner test for spurious correlation (see Law and Marks 2009, 360-361 n. 10). The results of these tests suggest that licensure was not correlated with other factors that independently affect minority labor force participation.

Evidence KPV claim we have ignored

KPV argue that there is much textual evidence that we have ignored that shows the discriminatory nature of licensing during the early twentieth century. We disagree. The fact that we may not have cited some of the particular sources quoted by KPV does not mean that we are unaware of that discrimination was widespread during this period and that licensing may have been viewed by some as a mechanism for reducing competition from minority workers. Indeed, in the introduction to our article, we wrote:

> Licensing laws may reduce the prevalence of minorities, either because minorities find it more costly to meet licensing requirements, or because licensing represents a deliberate effort to exclude minorities. While in the first instance a decline in minority representation is an unintended consequence of licensing, in the second, licensing allows regulatory authorities and incumbent practitioners to indulge in their taste for discrimination. (Law and Marks 2009, 352)

Indeed, KPV gloss over the fact that we do find econometric evidence that licensing may have harmed minority participation in some occupations. For instance, we find that the representation of blacks among barbers was significantly reduced by barber licensing laws. We argue that licensing was likely to harm blacks in barbering because the possession of a barbering license provides very little information about quality (reputation mechanisms should be sufficient) and because barbering was an occupation for which black workers were a potential competitive threat to white workers.[8] In a footnote (359 n. 8), we also mentioned textual evidence that suggests that licensing of barbers was adopted with discriminatory intent.

8. We also found that licensing of beauticians had a negative and statistically significant effect on black participation among beauticians if we exclude southern states from the regression. See Law and Marks (2009, 359 n. 9).

However, the fact that licensing was harmful to minorities in *some* occupations hardly establishes that licensing was harmful to minorities in *all* occupations. Additionally, our findings do not preclude discrimination against some minorities, even in those occupations where the net effect on minority participation was positive. After all, our estimates represent an *average treatment effect*. One of the advantages of an econometric/statistical approach is that it brings more systematic evidence to bear on an issue that can be gleaned from a survey of qualitative sources. Sometimes econometric findings are consistent with the qualitative evidence. Sometimes they are not. When they are not, the inclination is to put more weight on the econometric evidence, precisely because it is a more systematic approach. We are hardly unique in having this preference.

Another reason we place more emphasis on our statistical evidence is that it allows us to separate the intentions of the actors from the outcomes. Much of the qualitative evidence that KPV cite suggests that incumbent practitioners *desired* licensing in order to discriminate against black workers. KPV then prematurely jump to the conclusion that licensing must have had this effect. We do not dispute that there were white workers in early twentieth century America who wanted licensing to reduce competition from blacks. However, it does not necessarily follow that licensing, when adopted, actually *succeeded* in reducing competition from blacks. It is only by conducting an econometric analysis of licensure across a broad range of occupations that we can determine how licensure actually affected minorities.

In fact, it is this systematic, econometric approach that allows us ultimately to paint a far more historically nuanced portrait of the effects of licensure than KPV have acknowledged: namely, that the effects of licensing on minority participation during this time period depended on the extent to which minority workers were a competitive threat, as well as the degree to which there was concern about worker quality. In a few instances, licensing was indeed harmful to minorities, but in other instances, it was helpful. The advantage of our approach is that it allows us to compare objectively the actual effects of licensing across a wide range of occupations. We suspect it would be far more difficult to reach this conclusion by surveying what was written at the time about what individual actors or organizations wanted from licensing.

Theoretical debate about licensing

KPV's final complaint is that we ignore the critical debate about licensing and that we "act as though the quality-assurance rationale is something that the critical literature has neglected" (KPV 2012, 226). We have very little to say in

reply, largely because we do not believe that we have ignored the critical literature, nor do we claim to have invented the quality-assurance rationale for licensing. In our introduction we identified the two main theories for licensure (i.e., entry barriers vs. quality assurance) and cited the relevant sources. Our contribution here is empirical (i.e., to distinguish between the two main hypotheses that have been advanced to explain licensing), not theoretical. The only theoretical twist we may have added to the issue is to link the quality-assurance argument for licensing with a statistical discrimination story (the usual quality-assurance story does not tell us how licensing might affect *minority* workers, whereas combining quality assurance with statistical discrimination generates different predictions for minorities). However, we do believe that we have added to the empirical literature since, unlike most earlier studies, we are able to control for omitted time-invariant state-level factors that affect minority participation, and our study covers a broader range of occupations.

KPV seem to believe that we are uninformed about the literature on market solutions to asymmetric information problems. Obviously, we are aware of this literature. Indeed, because of the availability of these market-based mechanisms, we do not expect licensing to provide a quality assurance role in all instances (e.g., barbers). Accordingly, licensing is unlikely to be helpful (and may even be harmful) for minorities in environments where consumers can easily discover worker quality, either because quality is not an issue or because alternative quality-assurance mechanisms are available. The fact that we did not embark on a lengthy discussion of these alternative mechanisms does not mean that we are unaware of them.

Our problem with KPV's position is that it seems to be based on two non-trivial assumptions: the first is that if market solutions to asymmetric information problems *could* work, they *will* work, and therefore licensing is redundant. The second is that because licensure has shortcomings, it will necessarily fail. We do not dispute that there are plenty of good examples to support both assumptions in some cases. However, this is hardly a sufficient reason to assume that both assumptions are necessarily true in all instances. The fact that there may be potential market solutions does not mean that they will always work effectively, nor that licensing has nothing to add. Additionally, the fact that licensing may be an imperfect institution does not mean that its effects will be uniformly negative. Ultimately, it is an empirical issue. Our goal is simply to determine what the actual effects are without the imposition of strong priors.

Conclusion

No econometric study is perfect and few are definitive. As empirical social scientists, we must live with the fact that to address many questions, we must resort to observational data, which are recorded with error. Additionally, we have to accept that there are factors that may be relevant for our analysis but for which we cannot control. Finally, there are all the other human errors that arise, intentionally or otherwise, simply because none of us is infallible. In dismissing a particular empirical study, it is therefore not sufficient to show that a study is imperfect. Instead, one must go farther and demonstrate that the study is systematically biased in favor of what it finds.

KPV have leveled several criticisms against our study of the effects of occupational licensing laws on minorities during the Progressive Era. In our reply we have taken each of these criticisms seriously. While we acknowledge that there are some precise causal mechanisms that cannot easily be identified, and that in some instances we have erred in interpretation (certification vs. licensure), our view is that KPV have not successfully argued that our study is so biased that we make no contribution to the literature on how licensing laws affected minorities. They have failed to show that our use of census data, the uneven enforcement of licensing laws, measurement error in the timing and stringency of licensing laws, or the problem of omitted variables systematically bias our estimates in favor of finding positive effects. In each of these instances, we have shown that the problems that KPV have identified would either bias our estimates toward zero or in the opposite direction of what we do find. When we do estimate new regressions in light of KPV's suggestions, the results are consistent with our original findings. Specifically, we find that licensure increased representation of blacks among real estate agents and registered nurses, and increased representation of women in the real estate profession. Finally, KPV's falsification tests do not clearly demonstrate that our methodology is biased in favor of finding a positive effect on minority participation. Accordingly, KPV have not made a persuasive case that our study should be disregarded.

KPV also believe that we ignore the qualitative evidence in favor of discriminatory intent. While we dispute the claim that when it comes to our scholarly judgment, the "textual evidence of discriminatory intent…count[s] for very little" (KPV 2012, 226), our position is that it is impossible to know how licensing affected outcomes without a systematic, statistical analysis of the data. It is easy enough to find anecdotal evidence in favor of one position or another. The value added by modern empirical economics is that it goes beyond the anecdotes to

bring systematic data to bear on important economic questions. Accordingly, we are prepared to plead guilty to the charge of basing our conclusions on the statistical evidence. Would not any empirical economist plead the same way?

References

Acemoglu, Daron, and Jörn-Steffen Pischke. 2003. Minimum Wages and On-the-Job Training. *Research in Labor Economics* 22: 159-202

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. How Much Should We Trust Differences-In-Differences Estimates? *Quarterly Journal of Economics* 119: 249-275.

Council of State Governments. 1952. *Occupational Licensing Legislation in the States*. Chicago: Council of State Governments.

Federman, Maya N., David E. Harrington, and Kathy J. Krynski. 2006. The Impact of State Licensing Regulations on Low-Skilled Immigrants: The Case of Vietnamese Manicurists. *American Economic Review* 96: 237-241.

Hirsh, Barry T., David A. Macpherson, and Wayne G. Vroman. 2001. Estimates of Union Density by State. *Monthly Labor Review* 124: 51-55.

Klein, Daniel B., Benjamin Powell, and Evgeny S. Vorotnikov. 2012. Was Occupational Licensing Good for Minorities? A Critique of Marc Law and Mindy Marks. *Econ Journal Watch* 9(3): 210-233. **Link**

Kleiner, Morris M., and Alan B. Kreuger. 2009. Analyzing the Extent and Influence of Occupational Licensing on the Labor Market. *NBER Working Paper* No. 14979. National Bureau of Economic Research (Cambridge, Mass.).

Law, Marc T., and Sukkoo Kim. 2005. Specialization and Regulation: The Rise of Professionals and the Emergence of Occupational Licensing Regulation. *Journal of Economic History* 65: 723-756.

Law, Marc T., and Mindy S. Marks. 2009. Effects of Occupational Licensing Laws on Minorities: Evidence from the Progressive Era. *Journal of Law and Economics* 52: 351-366.

Law, Marc T., and Mindy S. Marks. 2012. Occupational Licensing and Minorities: A Reply to Vorotnikov, Powell, and Klein. April 19. Unpublished manuscript.

Monheit, Alan C. 1982. Occupational Licensure and the Utilization of Nursing Labor: An Economic Analysis. In *Advances in Health Economics and Health Services Research: A Research Annual*, vol. 3, ed. Richard M. Scheffler and Louis F. Rossiter, 117-142. Greenwich, Conn.: JAI Press.

Neumark, David, and William Wascher. 2001. Minimum Wages and Training Revisited. *Journal of Labor Economics* 19: 563-595.

Vorotnikov, Evgeny S., Benjamin Powell, and Daniel B. Klein. 2012. Was Occupational Licensing Good for Minorities? A Critique of Marc Law and Mindy Marks. March. Unpublished manuscript.

About the Authors

Marc T. Law is an associate professor in the Department of Economics at the University of Vermont. He is an applied microeconomist with research interests in the fields of regulation, political economy, and economic history. He has written extensively about the origins and effects of regulation in early twentieth-century America. His most recent project concerns the political economy of federalism and urban growth. His work has appeared in the *Journal of Law and Economics*, the *Journal of Law, Economics, and Organization*, the *Journal of Economic History*, the *Journal of Regional Science*, and other outlets. His email address is marc.law@uvm.edu.

Mindy Marks is an associate professor in the Department of Economics at the University of California-Riverside. She conducts research in applied microeconomics with an emphasis on labor, health, and education topics. Her projects to date involve large-scale empirical evaluations that use careful statistical analysis to determine underlying causal relationships. Her work has been published in the *Review of Economics and Statistics*, *Journal of Human Resources*, *Journal of Law and Economics*, and other outlets. She a founding co-editor of *Policy Matters*, a quarterly series of reports that provide timely research and guidance on issues that are of concern to policymakers at the local, state, and national levels. Her email address is mindy.marks@ucr.edu.

Go to Archive of Comments section
Go to September 2012 issue

EJW
ECON JOURNAL WATCH
Scholarly Comments on Academic Economics

ECON JOURNAL WATCH 9(3)
September 2012: 256-297

Ziliak and McCloskey's Criticisms of Significance Tests: An Assessment

Thomas Mayer[1]

LINK TO ABSTRACT

If economists have natural constants, then the most well-known is 0.05.
—Hugo Keuzenkamp and Jan Magnus (1995, 16)

Significance tests are standard operating procedures in empirical economics and other behavioral sciences. They are also widely used in medical research, where the prevalence of small samples makes them particularly welcome. And—mainly in the form of error bars—they are also at home in the physical sciences (see Horowitz 2004). But they have many critics, particularly among psychologists, who have done much more work on the topic than have economists.[2] Some members of the American Psychological Association even tried to ban their use in all journals published by the Association. That proposal was easily defeated, but some of the editors of those journals moved on their own to discourage significance tests, although with little effect; significance tests still reign in psychology. In medical research, however, the critics appear to have had considerable influence.[3]

1. University of California, Davis, Davis, CA 95616. I am indebted for excellent comments to Kevin Hoover and Deirdre McCloskey and to three outstanding referees for this journal. An earlier version was presented at the 2010 meeting of the Society of Government Economists.

2. For telling quotations from critics see Johnson (1999).

3. Fidler et al. (2004b, 626) explain the spread of the reform in part by a shift from testing to estimation that was facilitated by the medical literature, unlike psychology, using a common measurement scale, to "strictly enforced editorial policy, virtually simultaneous reforms in a number of leading journals, and the timely re-writing [of] textbooks to fit with policy recommendations." But their description of the process suggests that an accidental factor, the coincidence of several strong-willed editors, also mattered. For the classic collection of papers criticizing significance tests in psychology see Morrison and Hankel (1970), and for a more recent collection of papers see Harlow et al. (1997). Nickerson (2000) provides a comprehensive survey of this literature.

Within economics, although significance tests have been occasionally criticized for many years (see for instance White 1967 and Mayer 1980), these criticisms became prominent only in 1985 with the publication of D. N. McCloskey's *The Rhetoric of Economics* and a subsequent series of papers by McCloskey and by Stephen Ziliak that culminated in their 2008 book, *The Cult of Statistical Significance*.[4] There they charge: "Statistical significance is not the same thing as scientific finding. R^2, *t*-statistic, *p*-value, *F*-test, and all the more sophisticated versions of them in time series and the most advanced statistics are misleading at best. … [M]ost of the statistical work in economics, psychology, medicine, and the rest since the 1920s…has to be done over again" (Ziliak and McCloskey 2008b, xv and 18). They are "very willing to concede some minor role to even mindless significance testing in science" (48), but they declare: "Significance testing as used has no theoretical justification" (Ziliak and McCloskey 2004, 527).

Reception of Ziliak and McCloskey's book

The Cult of Statistical Significance has been widely, and in many cases very favorably, reviewed by journals in diverse fields, such as *Science*, *Administrative Science Quarterly*, *Contemporary Sociology*, *SIAM News*, *Nature*, *Medicine*, and *Notices of the American Mathematical Society*, as well as by some economics journals. Few, if any, books by contemporary economists have stirred interest in so many fields.

The main criticisms of Ziliak and McCloskey (henceforth referred to as Z-M) have come from papers in a symposium in the 2004 *Journal of Socio-Economics*, and from Kevin Hoover and Mark Siegler (2008a), Tom Engsted (2009), and Aris Spanos (2008). I will discuss most of the criticisms in the context of specific claims made by Z-M. But some of Spanos' criticisms are too general for that, since they raise fundamental and disturbing questions about the whole approach to significance testing (see Mayo 1996; Mayo and Spanos 2006; Spanos 2008). Thus, one disturbing problem is that significance testers pay insufficient attention to the problem of statistical inadequacy. If the model does not fit the data, e.g., if it is linear when the data embody a log-linear relationship, then the estimated standard errors and t-values are incorrect. Spanos calls this a problem for which no adequate solution has been discovered, and he guesses that, of the applied papers published in the *American Economic Review* over the last 30 years, less than one percent would pass a test of statistical adequacy (Spanos 2008, 163), thus implicitly agreeing with Z-M in rejecting published significance tests in economics.

4. Following the standard practice of focusing on an author's most recent statement of his or her thesis I will primarily discuss their 2008 book. There they take a more radical position than in their prior papers. As McCloskey (2008) explained, their frustration at having their more moderate statement ignored drove them towards making stronger statements.

Among Spanos' other criticisms are that Z-M play to the galleries, with repetitious and provocative catchphrases. One might reply that this makes for exciting reading, but it can also cover up inadequacies in the analysis. He also argues that Z-M's rejection of Deborah Mayo's (1996) interpretation of significance tests as measures, not of the probable truth of the hypothesis, but of the severity of the test, is based on a misunderstanding, and that their discussion of the relations among sample size, the power of a test, and the probability that the test will reject the null is confused.

In their reply Z-M (2008a, 166) claim—without offering any justification—that Spanos "throws up a lot of technical smoke" that hides his basic agreement with them, a claim that is hard to accept. They further state that the "*sole* problem" (166, italic in original) their book deals with is the distinction between statistical and substantive significance, and they blame Spanos for not appreciating that. There is something to this argument (even if their book might not leave the reader with that impression), in that this distinction, simple as it is, needed to be reiterated and that Spanos does not give them sufficient credit for doing so. Spanos, who approaches the matter wearing the spectacles of a mathematician and statistician, is concerned about logical problems with significance tests, such as the failure to ensure that the data conform to the assumptions underlying these tests, while Z-M, wearing the spectacles of applied economists, are concerned with invalid conclusions that appear in the *American Economic Review* (*AER*) and similar journals.

The focus of this paper

I will evaluate Z-M's claims only with respect to economics, even though this may be unfair to Z-M since their criticisms may be more applicable to other fields. (The subtitle of their book is: *How the Standard Error Costs Us Jobs, Justice and Lives*.) I will try to show that although their extreme claims are unwarranted, less extreme versions of some of their claims are correct. In doing so I take a pragmatic, second-best approach, looking only at those errors in the application of significance tests that are likely to cause readers to draw substantially wrong conclusions about substantive issues. As noted by Gerd Gigerenzer (2004) and Heiko Haller and Stefan Krauss (2002), careless statements about significance tests abound. In other words, my orientation is more akin to that of an engineer solving a practical problem in a rough and ready way, than that of a mathematician seeking elegant truths (see Colander 2011).

Since this is not a review of Z-M's book I also leave aside some topics that the book discusses at length, such as the history of significance tests. Their book is as much a history of significance tests as it is a discussion of their current use. For an evaluation that encompasses this and other items that I omit, see Spanos

(2008), who is very critical of Z-M's condemnation of R. A. Fisher. Similarly, I do not discuss a referee's charge that when Z-M complain about the confusion of statistical with substantive significance they merely reiterate what many psychologists and statisticians have said over several decades.

I therefore focus on the lessons that applied econometricians should garner from Z-M's work and from various responses to it. This debate has generated much heat and also an all-or-nothing stance on both sides. But with first-rate economists involved on both sides, it is unlikely that either side is totally wrong. I try to show that Z-M have made a substantial contribution, albeit not nearly as great a one as they seem to think. This should not be surprising; economics shows many valuable contributions, such as the permanent income theory, that were initially accompanied by excessive claims.

The emphasis of this paper is therefore more on an evaluative presentation of the debate than on entirely new evidence. Authors of most papers can assume that most of those who read their papers are specialists who are more or less familiar with the prior debate. But this is not the case here; those for whom the paper is relevant include nearly all readers of applied econometrics papers, and presumably most of them have not read Z-M and their critics.

The meaning of significance tests

Discussions of significance tests are not always clear about what they mean by the term. Some seem to mean any standardized statistical measure of whether certain results are sufficiently unlikely to be due to sampling error, including for example Neyman-Pearson methods, while others seem to define the term more narrowly as Fisherian tests. The broader definition seems more common in psychology than in economics. Z-M deal with both definitions, but reserve their special wrath for Fisherian tests.

The literature is also unclear about the source of the variance that underlies any significance test. There are three potential sources, measurement errors, sampling errors, and specification errors. Yet the literature sometimes reads as though the only problem is sampling error, so that with a 100 percent sample significance tests would be meaningless.[5] However, as Tom Engsted (2009) points out, economists generally do not aim for true models in the sense that the only deviation

5. Thus Z-M ridicule significance tests by pointing out that they are sometimes thoughtlessly used in cases where the sample comprises the entire universe, so that the notion of sampling error is inapplicable. In response, Hoover and Siegler (2008a) argue that although a paper may seem to include the entire universe within its sample, for instance the pegging of long-term interest rates in the U.S. after WW. II, such a paper is really intended to explain what happens *in general* when long-term interest rates are pegged. It is therefore using a sample. But Hoover and Siegler fail to note that it is not likely to be a random sample for all cases when interest rates are pegged, and to that extent standard significance tests are not applicable. However, to the extent that variations between individual observations are due to random measurement errors, t-values regain their meaning.

between their predictions and the data are sampling errors, but for models that are useful. And for that it does not matter if these models result in systematic rather than unsystematic errors. Hence, argues Engsted (2009, 394, italics in original), economists "to an increasing degree hold the view…that we should *not* expect model errors to be unsystematic…. Such models will be statistically rejected at a given significance level if the test is sufficiently powerful. Economists, therefore, to an increasing extent…evaluate economic models empirically using methods that are better suited for misspecified models than [are] statistical hypothesis tests." Specific examples he cites include dynamic stochastic general equilibrium (DSGE) models, linear rational expectations models, and asset pricing models. Spanos (2008) and Walter Krämer (2011), too, argue that Z-M do not treat the problem created by specification errors adequately.

Z-M's criticisms of significance tests

Z-M challenge economists' use of significance tests on four grounds. First, they claim that most economists do not realize that substantive significance, that is, the size of the effect that an independent variable has on the dependent variable (which Z-M call "oomph"), is vastly more important than statistical significance, or what is even worse, economists confound the two. Second, economists often commit the logical fallacy of the transposed conditional; third, they ignore the loss function; and fourth, instead of reporting confidence intervals, they usually present their results in terms of t-values, *p*'s, or F-ratios.

What matters, significance or oomph?

There are several issues under this rubric. One is: What it is that significance tests do? The second is whether the mere existence of an effect—as distinct from its size—is a legitimate scientific question. The third is the frequency with which economists and others focus on statistical instead of on substantive significance. A fourth, raised by a referee of this paper, is that we do not know how to measure oomph properly, that a "point estimator is clearly not a good measure of 'oomph'" and that "nobody knew how to address the problem!" (Anonymous 2012). But the world's work has to be done, and I will therefore arbitrarily assume that the point estimate is a sufficiently good measure of oomph, while admitting that we have here another reason for being modest in our claims.

What do significance tests tell us?

In at least some of their writings Z-M give the impression that significance tests only tell us (directly in the case of confidence intervals, and indirectly for t-values and *p*'s) the spread around a point estimate (or a sample mean) whose correct value we already know (see Horowitz 2004, 552). They suggest the following mental experiment:

> Suppose you want to help your mother lose weight and are considering two diet pills with identical prices and side effects. You are determined to choose one of the two pills for her. The first pill, named Oomph, will on average take off twenty pounds. But it is very uncertain in its effects—at plus or minus ten pounds. … Oomph gives a big effect, you see, but with a high variance. Alternatively the pill Precision will take off five pounds on average. But it is much more certain in its effects. Choosing Precision entails a probable error of plus or minus a mere one-half pound. … So which pill for Mother, whose goal is to lose weight? (Z-M, 2008b, 23)

This mental experiment is flawed (Hoover and Siegler 2008b, 15-16; Engsted 2009, 400.) It assumes that we already know the means for the two pills, and it thereby bypasses the need to ask the very question that significance tests address: Given the existence of sampling error, how confident can we be about these point estimates? Suppose your sample consists of only two cases for each pill. Shouldn't you then warn mother not to place much importance on what you told her about the two means? (See Wooldridge 2004.) But sample size alone does not tell her how much credence to give your means; variance also matters. So why not couch your warning in terms that combine sample size and variance, such as t-values, *p*'s or confidence intervals? It is Z-M's unwarranted assumption that we know the mean of the universe, rather than just the mean of the sample, that allows them to dismiss significance tests as essentially useless.

This failure to acknowledge that statistical significance is often needed to validate a paper's conclusions about oomph underlies Z-M's rejection of testing for both statistical significance as well as for oomph. Thus they write: "Statistical significance is *not* necessary for a coefficient to have substantive significance and therefore *cannot* be a suitable prescreen" (Z-M 2008b, 86, italics in original). Yes, statistical significance is not necessary for substantive significance, but that is not the issue. Statistical significance is needed to justify treating the coefficient generated by the sample as though it had been generated by the universe, i.e., as a sufficiently reliable stand-in for the true coefficient.

Part of Z-M's argument against the importance of statistical significance is couched as an attack on what they call "sign econometrics" and "asterisk econometrics", by which they mean placing asterisks next to coefficients that are significant and have the right sign, because researchers mistakenly believe that what makes a variable important is its significance along with the right sign, and not its oomph.[6] But this is not necessarily so. Putting asterisks on significant variables does not necessarily imply that they are more important than others; the importance of a variable can be discussed elsewhere. And readers should be told for which coefficients there is a high likelihood that their difference from zero is not just the result of sampling error. Mayo (1996) provides a credible argument that one should interpret a t-value, not as an attribute of the hypothesis being tested, but as an attribute of the severity of the test to which it has been subjected. And there is nothing wrong with using asterisks to draw attention to those hypotheses that have passed a severe test.[7]

Is existence a scientific question?

There is also a flaw in Z-M's claim that significance tests only tell us whether an effect exists, and that this is a philosophical and not a scientific question. But existence *is* a scientific question, in part because it may tell us whether we are using the right model (see Elliott and Granger 2004) and because it makes little sense for scientists to try to measure the size of something that does not exist. It would be hard to obtain an NSF grant to measure the density of ether. And we do observe natural scientists asking about existence.[8] Hoover and Siegler (2008b) cite a classic test of relativity theory, the bending of light near the sun, as an example where oomph is irrelevant while significance is crucial.[9] Recently there was much excitement about neutrinos allegedly traveling faster than light because relativity theory prohibits that, even if it is only trivially faster. Wainer (1999) lists three other examples from the natural sciences where oomph is not needed: (a) the speed of

6. Z-M do, however, allow for exceptions to their condemnation, writing: "*Ordinarily* sign alone is not *economically* significant unless the magnitude attached to the sign is large or small enough to matter" (2008b, 70, first italic added, second in original).

7. By "severity" I mean the probability that the test would reject the null even if it were true.

8. Thus Hoover and Siegler (2008a, 27) show that, contrary to Z-M's claim, physical scientists also use significance tests. In their reply McCloskey and Ziliak (2008, 51-52) concede this, but surmise that they do so much less frequently than economists do. However, *if* physical scientists typically have larger samples than economists (perhaps because they can rerun their experiments many times) they might have less need for significance tests (see Stekler 2007).

9. Z-M (2008b, 48-49), however, reject this interpretation of that test because statistical significance did not play a role in it. But the issue here is whether existence matters in science, and not whether significance tests happen to be used to establish existence.

light is the same at points moving at different speeds; (b) the universe is expanding; (c) the distance between New York and Tokyo is constant. Horowitz (2004) cites additional examples from physics.)

Within economics, Giffen goods provide an example. Economists are interested in knowing whether such goods exist, regardless of the oomph of their coefficients. In Granger tests, too, what counts is the existence of an effect, not its oomph (Hoover and Siegler 2008a). And that is so also when we use a likelihood ratio to choose between a restricted and a more general model (see Robinson and Wainer 2002). Even when we are interested in oomph, it does not always matter more than existence does. Consider the hiring policy of a firm. Suppose a variable measuring race has a much smaller oomph in a regression explaining the firm's employment decisions than an education variable does. As long as the racial variable has the expected sign and is significant, you have bolstered a claim of racial discrimination. By contrast, suppose you find a substantial oomph for the racial variable, but its *t* is only 1.0. Then you do not have as strong a case to take to court. Z-M might object that this merely shows that courts allow themselves to be tricked by significance tests, but don't courts have to consider some probability of error in rendering a verdict?

One can go beyond such individual cases by dividing hypotheses into two classes. One consists of hypotheses that explain the causes of observed events. For these oomph is generally important. If we want to know what causes inflation, citing declines in strawberry harvests due to droughts will not do, even if, because of the great size of the sample, this variable has a t-value of 2. But there is also another type of model (and for present purposes one need not distinguish between models and hypotheses), one that Allan Gibbard and Hal Varian (1978) called a "caricature model", which tries to bring out important aspects of the economy that have not received enough attention. And these aspects may be important for the insight that they provide (and hence for the development of new causally-oriented hypotheses), even though the variables that represent these aspects have little oomph in a regression equation.

For cause-oriented hypotheses, regression tests can be categorized as direct or indirect. Direct tests ask about whether the variable has a large oomph, or if it explains much of the behavior of the dependent variable. If we find neither oomph nor statistical significance we consider the hypothesis unsatisfactory. But we frequently also use indirect tests. These are tests that draw some necessary implication from the hypothesis and test that, even though this particular implication is of no interest on its own. Here oomph does not matter, but the sign and significance of the coefficient do. The additional opportunities for testing that these indirect tests provide are important parts of our toolkit because we frequently lack adequate data for a direct test.

For instance, Milton Friedman (1957) encountered a problem in testing the permanent income theory directly because no data on permanent income were then available. He therefore drew implications from the theory, for example, that at any given income level farm families have a lower marginal propensity to consume than urban families, and tested these implications (see Zellner 2004). Surely, few readers of Friedman's *A Theory of the Consumption Function* found the relative size of these propensities to consume interesting for their own sake, and therefore did not have any interest in their oomph, but the sign of the difference and its significance told them something about the validity of the permanent income theory. Friedman presented eight tests of this theory. Seven are indirect tests.[10] Standing by itself, each of these seven tests is only a soft test because it addresses only the direction of a difference, and a priori, without the availability of the theory, there is a 50 percent chance that the difference will be in the predicted direction. But if all of these seven tests yield results in the direction predicted by the permanent income theory, then one can invoke the no-miracles argument.

Or suppose you test the hypothesis that drug addiction is rational by inferring that if the expected future price of drugs rises, current drug consumption falls. And you find that it does. To make sure that you have a point, you need to check whether this reduction is large enough so as not to be attributable to sampling error. But it does not have to account for a substantial decline in drug use. This does not mean that oomph never matters for indirect tests; in *some* cases it may.

There are also in-between cases where oomph matters for some purposes but not for others. Take the standard theory of the term structure of interest rates. It seems logically compelling, but it does not predict future changes in the term structure well. If someone, by adding an additional variable develops a variant that does predict well, this will be of interest to many economists, both to those who want to predict future rates and to those who wonder why the standard theory predicts badly, regardless of the oomph of the new variable. Similarly, it would be useful to have a variant of the Fisher relation that predicts movements of exchange rates better, even if it means adding a variable with a low oomph.

Finally, the role of the coefficient's sign and its significance level are enhanced when one considers papers not in isolation, but as part of an ongoing discussion where the purpose of a paper may be to counteract a previous paper. For that it may suffice to show that in the previous paper, once one corrects for some error or uses a larger sample, the crucial coefficient has the wrong sign, or

10. And the one direct test Friedman provided did not support his theory against the rival relative income theory of Duesenberry and Modigliani. Hence, by concluding that his evidence supported the permanent income theory Friedman put more weight on the indirect tests than on the direct test. For a detailed discussion of Friedman's tests see Mayer (1972).

loses significance, and never mind its oomph. For example, it was widely believed that, prior to the Glass-Steagall Act, banks that underwrote securities had a conflict of interest that allowed them to exploit the ignorance of investors. The evidence cited was that in the 1930s securities underwritten by banks performed worse than others. By showing that at the five-percent significance level the opposite was the case, Randall Kroszner and Raghuram Rajan (1994) challenged this hypothesis without having to discuss oomph.

None of the above denies that in many cases oomph is central. But it does mean that Z-M's claim that existence is generally unimportant, while oomph is generally important, is invalid. As the first column of the Table based on a sample of 50 papers in the *AER* shows, oomph was neither required or very important in at least eight (16 percent) and arguably in as many as 12 (24%) of the papers. But while this result rejects Z-M's strong claim, it still means that, at least in economics, the oomph of strategic coefficients usually deserves substantial emphasis.

How often do economists confuse statistical significance with substantive significance?

Very often, say Z-M. Having surveyed 369 full-length *AER* articles from January 1980 to December 1999 that contain significance tests, they claim: "Seventy percent of the articles in… the 1980s made no distinction at all between statistical significance and economic or policy significance… Of the 187 relevant articles published in the 1990s, 79 percent mistook statistically significant coefficients for economically significant coefficients." (Z-M 2008b, 74, 80).[11] Even though there are many cases, such as reduced form usage of vector-autoregressions (VARs), where the magnitude of a particular coefficient is of little interest (Engsted 2009, 400), Z-M's results are still surprising, particularly since plotting confidence intervals is the default setting of frequently used econometric software packages (Hoover and Siegler 2008a, 20). Moreover, Engsted (2009) points to several flourishing areas of economic research, such as DSGE models and return predictability, where statistical and substantive significance are clearly *not* confounded.

How then did Z-M obtain their dramatic results? They did so by giving each paper a numerical score depending on its performance on a set of nineteen

11. Altman (2004, 523) in a less quantitative way supports Z-M's claim of frequent confusion, writing: "As both a researcher and a journal editor, I have been struck by the insistence of [*sic*] the use of tests of statistical significance as proxies for analytic significance. More often then not the writers are not aware of the distinction between statistical and analytic significance or do not naturally think of the latter as being of primary interest or importance."

questions, e.g., whether the paper refrains from using the term "significant" in an ambiguous way, and whether in the conclusion section it keeps statistical and economic significance separated (Z-M 2008b, 72-73). Such grading requires judgment calls. What is ambiguous to one reader may be unambiguous to another. Moreover, suppose a paper uses "significant" in an ambiguous way in its introduction, but in the concluding section uses it in a clear way. Should it be docked for the initial ambiguity? And what grade does a paper deserve that does not distinguish between statistical and economic significance in the conclusion, but does so at length elsewhere? It is therefore not surprising that Hoover and Siegler (2008a, 5) criticize Z-M for frequently making dubious judgments as well as for using a hodge-podge of questions, including some questions that indicate good practice, some that indicate bad practice, and some that are redundant. Moreover, as Hoover and Siegler point out, some questions duplicate others, which results in double counting and hence an arbitrary weighting.[12] And Hoover and Siegler found entirely unconvincing the five case studies that Z-M provide. Similarly, Jeffrey Wooldridge (2004, 578) wrote: "I think [Z-M] oversell their case. Part of the problem is trying to make scientific an evaluation process that is inherently subjective. It is too easy to pull isolated sentences from a paper that seem to violate ZM's standards, but which make perfect sense in the broader context of the paper." Spanos (2008, 155) calls "most" of Z-M's questions "highly problematic". Appendix A lists and evaluates each of Z-M's questions.

An alternative procedure is not to use predetermined questions to assess specific sentences in a paper, but to look at a paper's overall message and ask whether it is polluted by a failure to distinguish statistical from substantive significance. Anthony O'Brien (2004) selected a sample of papers published in the *Journal of Economic History* and in *Explorations in Economic History* in 1992 and 1996 and asked whether their conclusions were affected by an inappropriate use of significance tests. In 23 out of the 118 papers (19 percent), significance tests were used inappropriately, but in only eight of them (7%) did "it matter for the paper's main conclusions" (O'Brien 2004, 568). This low percentage, he suggested, may explain why Z-M have had so little success in moving economists away from significance tests.

In my own attempt to replicate Z-M's results I followed O'Brien, as did Hoover and Siegler, by looking not at the specific wording of particular sentences but at a paper's overall Gestalt. I used a sample of 50 papers, thirty-five of them taken from Z-M's sample (17 from the 1980s and 18 from the 1990s), and, to update the sample, fifteen papers from 2010. Specifically, I asked whether a harried

12. For Z-M's reply and Hoover and Siegler's rejoinder, see McCloskey and Ziliak (2008) and Hoover and Siegler (2008b).

reader not watching for the particulars of significance testing would obtain the correct takeaway point with respect to significance and oomph.[13] Admittedly, this procedure requires judgment calls, and other economists may evaluate some papers differently from the way I do.[14] But Z-M's criteria also require judgment calls. So that readers can readily judge for themselves which procedure is preferable, Appendix B provides summaries of the eleven papers in my sample that Z-M give a low grade.

The second column of the Table shows the results. A "yes" means that the authors do give the oomph. Since a yes/no dichotomy often does not capture the subtlety of a discussion, dashes and footnotes indicate in-between cases. The Table treats as a "yes" cases where the authors do not discuss oomph in the text but do give the relevant coefficients in a table. It may seem necessary to discuss oomph in the text and not just in a table, because that allows the author to tell readers whether the coefficient should be considered large or small, something that may not always be obvious, particularly if the regression is in natural numbers rather than logs. For example, if we are told that by changing the bill rate by ten basis points, the Fed can change the five-year rate by one basis point, does this mean that it has sufficient or insufficient control over the latter? But even in a brief discussion in the text it may sometimes be hard to say that a coefficient is "large" or "small", because the results of the paper may be relevant for several issues, and what is a large and important oomph with respect to one issue may not be so for another.[15] And even for the same issue it may vary from time to time: When the bill rate is five percent it does not hinder the Fed as much if it requires a change in the bill rate of 30 basis points to change the five-year rate by 10 basis points as it does when the bill rate is 0.25 percent. A requirement that oomph be discussed in the text would therefore not be a meaningful criterion by which to distinguish between those who use significance tests correctly and those who don't. I have therefore used a minimal criterion, that the coefficient be given, so that readers can make up their own minds.

13. I assume a harried reader because, given the great volume of reading material that descends upon us, it seems unlikely that most papers receive a painstaking reading. Further, I focus on a paper's main thesis, and therefore do not penalize it if it fails to discuss the oomph of a particular variable that is not strategic, even if this oomph is interesting for its own sake.

14. In fact, in some cases I changed my mind when I reviewed an earlier draft.

15. Moreover, even with respect to any one issue the magnitude of oomph may not answer the question of interest, because (leaving the causality issue aside) all it tells you is by how much y changes when x changes by one unit. It does *not* tell you what proportion of the observed changes in y is due to changes in x, because that depends also on the variance of x—a statistic that should be given, but often is not.

TABLE. Uses and misuses of significance tests in 50 *AER* papers

Papers:	(1) Oomph required or very important for topic	(2) Paper provides oomph	(3) Paper provides correct takeaway point with respect to oomph	(4) Significance tests used wrong-way-round for testing maintained hypothesis	(5) Significance tests used wrong-way-round for congruity adjustments[a]
1980s:					
Acs & Audretsch (1988)	Yes	Yes	Yes	Yes	No
Blanchard (1989)	Yes	Yes	Yes	No	--[b]
Bloom & Cavanagh (1986)	Yes	Yes	Yes	No	Yes
Borjas (1987)	Yes	Yes	Yes	--[c]	No
Carmichael & Stebbing (1983)	Yes	Yes	Yes	No	Yes[d]
Darby (1982)	Yes	Yes	Yes	No	No[e]
Evans & Heckman (1984)	No	No	Irrelevant	No	No
Froyen & Waud (1980)	No	Yes	Yes	No	No
Garber (1986)	Yes	Yes	Yes	Yes	No
Johnson & Skinner (1986)	Yes	Yes	Yes	No	No
Joskow (1987)	Yes	Yes	Yes	--[f]	No
LaLonde (1986)	Yes	Yes	Yes	No	No
Mishkin (1982)	No[g]	Yes	Yes	--[h]	Yes
Pashigian (1988)	--[i]	Yes	--[j]	Yes	No
Romer (1986)	No	Yes	Yes	Yes	No
Sachs (1980)	Yes	Yes	Yes	No	No
Woodbury & Spiegelman (1987)	Yes	Yes	Yes	No	No
1990s:					
Alesina & Perotti (1997)	Yes	Yes	Yes	No	No
Angrist & Evans (1998)	Yes	Yes	Yes	No	No
Ayres & Siegelman (1995)	Yes	Yes	Yes	No	No
Borjas (1995)	Yes	Yes	Yes	No	No
Brainard (1997)	Yes	Yes[k]	Yes	No	No
Feenstra (1994)	Yes	Yes	Yes	--[d]	No
Forsythe (1992)	No	Yes	Yes	Yes	No
Fuhrer & Moore (1995)	Yes	Yes	Yes	No	Yes[d]
Gali (1999)	No[m]	Yes	Yes	No	Yes
Ham et al. (1998)	Yes	Yes[n]	Yes	No	No
Hendricks & Porter (1996)	Yes	Yes	Yes	No	No
Hoover & Sheffrin (1992)	No	Yes	Irrelevant	Yes	No
Kroszner & Rajan (1994)	No	Yes	Yes	Yes	No
Mendelsohn et al. (1994)	Yes	Yes	Yes	No	No
Pontiff (1997)	Yes	Yes	Yes	No	No
Sauer & Leffler (1990)	No	Yes[n]	Yes	No	No
Trejo (1991)	Yes	Yes	Yes	No	No
Wolff (1991)	Yes	Yes	--[n]	No	No

Papers:	(1) Oomph required or very important for topic	(2) Paper provides oomph	(3) Paper provides correct takeaway point with respect to oomph	(4) Significance tests used wrong-way-round for testing maintained hypothesis	(5) Significance tests used wrong-way-round for congruity adjustments[a]
2010:					
Artuç et al. (2010)	Yes	Yes	Yes	No	No
Bailey (2010)	Yes	Yes	Yes	Yes	No
Bardhan & Mookherjee (2010)	No	Yes	Yes	No	Yes
Chandra et al. (2010)	Yes	Yes	Yes	No	No
Chen et al. (2010)	Yes	Yes	Yes	No	No
Conley & Udry (2010)	Yes	Yes	Yes	No	No
Dafny (2010)	No[o]	Yes	Yes	No	No
Ellison et al. (2010)	Yes	Yes	Yes	Yes	No
Fowlie (2010)	Yes	--[p]	Yes	No	No
Harrison & Scorse (2010)	Yes	Yes	Yes	No	No
Landry et al. (2010)	Yes	Yes	Yes	No	No
Lerner & Malmendier (2010)	Yes	Yes	Yes	--[q]	No
Leth-Petersen (2010)	Yes	Yes	Yes	No	No
Mian et al. (2010)	Yes	Yes	Yes	Yes	No
Romer & Romer (2010)	Yes	Yes	Yes	No	No

Notes:
a. Includes tests for breaks in series, such as tests for unit roots, lag length, breaks in time series, etc. As discussed in the text, the frequency with which decisions about data adjustments have been made on the basis of wrong-way-round significance tests is probably understated.
b. Blanchard uses a wrong-way-round test in defending his assumption of stationary, but this is mitigated by his openness about the problem and his stating that theory suggests stationarity, as well as his saying: "as is…well known, the data cannot reject other null hypotheses. … [T]he results…must be seen as dependent on a priori assumptions on the time-series properties of the series" (Blanchard 1989, 1151). I believe that this absolves Blanchard of the charge of misusing the significance test.
c. Not clear how serious the problem is here.
d. Only at an unimportant point.
e. Arguably "yes" as Darby uses results from another paper in which insignificant variables were eliminated.
f. Wrong-way-round significance test used only at a minor point.
g. Oomph not required because the paper obtains negative results for the hypothesis it tests.
h. Wrong-way-round significance test used in auxiliary, but not in the main regressions.
I. Not really needed, but would be useful.
j. Provides oomph, but at one point makes the error of interpreting the size of a regression coefficient as a measure of the extent to which this regressor accounts for changes in the dependent variable; that depends also on the variance of that variable and on the variance of the dependent variable.
k. Provides oomph sometimes, but not frequently enough.
l. Main message of paper is qualitative.
m. Required only for comparison with other hypotheses that are mentioned only briefly.
n. On most, but not all points.
o. Only to the extent that oomph is not trivial.
p. Mainly, but not completely.
q. At one point uses what seems like a wrong-way-round test, but that is ameliorated by using it only to state that this test "provides no evidence" for a difference between two coefficients.

The results shown in the second column of the Table are in sharp contrast to Z-M's. There is no case where oomph is required but is totally ignored, and in only four cases might one reasonably say that it should have been given more emphasis. Further (though this is not shown in the Table) there is no evidence anywhere of a confusion of statistical significance and magnitude. As column (3) shows, for none of the fifty papers is their takeaway point *unequivocally* wrong due to the authors having confused statistical significance with oomph, or having failed to notice the importance of oomph, and in only two cases are the proffered takeaway points *arguably* wrong.

Krämer (2011, 9) examined all empirical papers in *The German Economic Review* since its inauguration in 2000. He found in 56 percent of them: "Confusion of economic and statistical significance of estimated coefficients or effects ('significant' used for both? [or] [m]uch ado about statistically significant but economically small coefficients or effects?)" In addition, 28 percent of the papers discarded (wrongly, he believes) "economically significant and plausible effects…due to lack of statistical significance." These results seem more discouraging than mine, but that may be explained by his criteria being more stringent. My tabulation does not penalize a paper for using the term "significant" for both types of significance. Nor does it ring an alarm bell when a variable with a substantively large oomph is disregarded because its t-value is less than 2.

Neither O'Brien's results nor mine (and certainly not Krämer's) should be read as a wholesale rejection of Z-M's contention. One reason is that in papers in some lesser-ranked journals—as well as in journals outside of economics—the confusion of statistical with substantive significance could well be more common.[16] Second, even if this confusion occurs only occasionally, that is too much. Given the vast number of papers that use significance tests, even a one percent error rate would mean many errors. Hoover and Siegler (2008a) criticize Z-M for making an already well known point in fussing about the distinction between statistical and substantive significance. They are right that the point is well known in an abstract sense, but if it is ignored, even only occasionally in actual practice, it is a point well worth making. Moreover, the error rate should be zero, since the distinction between the two types of significance is an elementary issue.

16. Mayo and Spanos (2006, 341) call the confusion between statistical and substantive significance the "[p]erhaps most often heard and best known fallacy" in using significance tests.

Wrong-way-round significance tests

Suppose you test your hypothesis that large banks benefit from scale economies. Not quite, says your computer, the relevant coefficient has the right sign, but a t-value of only 1.2. Can you now publish a paper showing that there are *no* scale economies for large banks? Only, Z-M tell us, if editors and referees are not doing their job. (See also Cohen 1994; Krämer 2011; Mayer 1980, 1993, 2001; Mayo 1996.) Except in the case of very large samples (discussed below), treating a low t-value as evidence that some effect does *not* exist is an error; Z-M refer to it as "the error of the transposed conditional."[17] Jacob Cohen (1994, 998, italics added) illustrates this error by contrasting the following two syllogisms:

> If the null hypothesis is correct, then this datum (D) can not occur.
> It has, however, occurred.
> Therefore, the null hypothesis is false.

And

> If H_0 is true this result (statistical significance) would *probably* not occur.
> This result has occurred.
> Then H_0 is probably not true and therefore formally invalid.

The second syllogism, unlike the first, is errant. For the first, since the premises are true the conclusion is also true. But in the stochastic world of the second we cannot be sure of the conclusion. To argue from the failure to disconfirm to the probability of confirmation one needs to look at the power of the test, that is, at the probability that if the hypothesis were true the test would not have disconfirmed it (see Mayo and Spanos 2006). But tests of power are unusual in economics.

An intuitively simple way of showing why one should not treat failure to disconfirm a hypothesis as confirmation of its converse is to point out that the results obtained when testing a null hypothesis fall into one of three bins: "confirmed", "disconfirmed", and "cannot tell". If the coefficient of x has a t-value of, say, 1.2, all this means is that the null hypothesis that x has no predictive value for y cannot be placed into the confirmed bin, but it does not authorize us to place it into

17. For a discussion of the transposed conditional from a Bayesian viewpoint see Cohen (1994), and from a frequentist viewpoint see Mayo and Spanos (2006). For a comprehensive survey of this problem in the psychological literature see Nickerson (2000), who describes it as an example of the logical fallacy of affirming the consequent.

the disconfirmed bin. If it did, it would be easy to disconfirm any hypothesis—just use a small enough sample.[18]

Sample size plays a fundamental role here, such a fundamental role that it might be useful to call these tests "sample-size" tests instead of significance tests, which would reduce the temptation to use them the wrong way round. As Edward Leamer (quoted in Leamer 2004, 557) has remarked: "Meaningful hypothesis testing requires the significance level to be a decreasing function of sample size." Just what does a significance test reject when $t < 2$? Is it the claim that the coefficient of the regressor exceeds zero for reasons other than sampling error, or is it the adequacy of the sample? (See also Altman 1980.)

And herein lies the kernel of validity in the use of wrong-way-round significance tests. Suppose, working with a sample of 100,000, one finds that a variable that the hypothesis predicts to be positive is positive but not significant. With such a large sample we have a strong expectation that a coefficient that is significant in the universe is also significant in our sample, so its insignificance speaks against the hypothesis. A power test, if available, would help in deciding. If not, we have to make a subjective judgment.

Two further arguments against wrong-way-round significance tests

My discussion seems to contradict the familiar Popperian principle that, due to the problem of induction, data can never prove a hypothesis but can only fail to disconfirm it. And if, again and again, independent and severe tests fail to disconfirm it, that justifies tentatively accepting it, at least as a working hypothesis.[19] This may seem to imply that when in a series of independent severe tests the beta coefficients of the relevant variable all have low t-values, we can treat the hypothesis as disconfirmed. But when philosophers speak of "failure to disconfirm" they mean failure to provide *any* evidence against the hypothesis. And

18. Moreover, the assumption that failure to disconfirm implies that the converse has been confirmed has an unwelcome implication. Suppose an economist tests the hypothesis that $y = x$ and finds that, though in his data set $x = 5$ and $y = 6$, he cannot reject at the five-percent level the null hypothesis that this difference is due only to sampling error. He therefore concludes that the data do not disconfirm his hypothesis, and that this increases its plausibility. His sister tests the contrary hypothesis that $y < x$, and since she uses the same data also finds that $x = 5$ and $y = 6$. Since in her test the difference between the predicted and the actual coefficients is again not significant, she, too, claims that the data confirm her hypothesis. Who is right?

19. For this the tests have to be hard, and not in Blaug's (1989, 256) classic description of much econometric testing, as "playing tennis with the net down", so that it would take a highly implausible combination of circumstances for a false hypothesis to have passed all of these tests. As Hoover (2011) has suggested the null hypothesis is therefore often a poor foil to the maintained hypothesis.

even if a coefficient with the right sign is significant only at, say, the 40 percent level, it still provides *some* evidence in favor of—and not against—the hypothesis that the coefficient is positive. In Mayo's formulation, it has failed a severe test, but it *has* passed a less severe test.

Congruity adjustments

My own survey (see the Table) distinguishes between significance tests that deal directly with a maintained substantive hypothesis and those that deal with whether a hypothesis needs certain adjustments—which I will call "congruity adjustments"—to make it congruent to the probability model that generated the data. For example, the data might have a log normal distribution while the hypothesis was initially formulated in terms of natural numbers. Other examples include tests for unit roots, serial correlation, heteroscedasticity, and breaks in the data.[20] Standard regression packages provide options for such adjustments, but their appropriateness in terms of the underlying hypothesis needs to be considered. The prevailing procedure is to make congruity adjustments only if one can reject at the five-percent level the null hypothesis that no adjustment is needed. But it is hard to see why the burden of the proof is thus placed on the hypothesis that an adjustment is needed. Brandishing Occam's razor will not do because not adjusting is only computationally and not philosophically simpler than adjusting. What we need, but do not have, is an explicit loss function. In testing the maintained hypothesis the usual justification for the five-percent level is that a Type II error is more damaging to science than a Type I error. But is this the case when one decides whether to adjust for, say, serial correlation? *Perhaps* a *p* value of 0.50 would be more appropriate. Unless we can decide on the appropriate loss function we should, when feasible, require our results to be robust with respect to these potential adjustments.

Frequency of wrong-way-round tests

As column (4) of the Table shows, even if one leaves congruity adjustments aside and looks only at tests of substantive hypotheses, 10 papers (20 percent) fall into the trap of assuming that failure to confirm a hypothesis at the five-percent level is equivalent to treating its negation as confirmed. And, as discussed in the Table notes, there are five additional papers that fall into the trap if one applies a stricter standard than I did, giving a potential total of 30 percent. Previously (Mayer

20. The same problem arises when deciding on the appropriate lag length by truncating when lagged coefficients become insignificant.

2001), I looked at papers in all the 1999 and 2000 issues of the *American Economic Review* and the *Review of Economics and Statistics* and found six cases of this confusion. The problem is worse in political science. There, Jeff Gill (1999) found that in four leading journals significance tests were used the wrong way round in 40 to 51 percent of the relevant cases.

The last column of the Table, which deals with congruity adjustment, shows four or at most six papers suffering from this error. But this underestimates—probably very substantially—the actual number of cases because it includes only those in which authors discuss their congruity adjustments. Presumably in many more papers authors tested for serial correlation, etc., and decided not to make the adjustment because it was not required at the five-percent level.

Permissible uses of wrong-way-round significance tests

However, none of the above implies that—even when the sample is not very large—one can never even tentatively reject a hypothesis because of a low t-value. Suppose $p = 0.15$. One can then consider a reasonable minimal value for the coefficient that would support the hypothesis, estimate the p for that, and from that decide on the credibility of the hypothesis (see Berg 2004; Mayo and Spanos 2006). For example, take the hypothesis that illegal immigration has lowered real wages in a certain industry. If this cannot be rejected at the five-percent level, one can test the hypothesis that it has reduced wages by more than a trivial two percent. Another possibility is to rely on a combination of tests. If on many independent tests a coefficient has the right sign but is not significant, one can either formally or informally reject the hypothesis that it is only sampling error that gives the coefficient the right sign (see Pollard and Richardson 1987). It is along these lines that Jean Perrin confirmed Einstein's interpretation of Brownian motion (see Mayo 1996, chapter 7).

The loss function

Although most economists think of significance tests as telling us something about a hypothesis, Z-M view it in a Neyman-Pearson framework, as telling us whether a certain course of action is justified. That requires a loss function.[21] The

21. Z-M tell us: "[W]ithout a loss function a test of statistical significance is meaningless. … [E]very inference drawn from a test of statistical significance is a 'decision' involving substantive loss… Accepting or rejecting a test of significance without considering the potential losses from the available courses of action…is not ethically or economically defensible." (2008b, 8-9, 15)

main issue here is by whom and at what stage it should be introduced, specifically whether it should be considered as part of the significance test, or instead as part of a larger problem for which we will use the results of the significance test as just one of several considerations. There is also the question of how to interpret a loss function in some of the situations in which we just want to satisfy our curiosity, and not to appraise some policy.

The conventional view is that the econometrician should deal with positivistic issues and turn the results over to the policymaker, who consults a loss function in deciding what action to take.[22] This has two putative advantages. First, it places most value judgments outside of economics. Second, as Hoover and Siegler (2008a) remind us, it avoids the problem that an econometrician's results may be relevant for many different policies, each of which calls for its own value judgments and hence has its own loss function. How is the econometrician to know all these loss functions, particularly when some of the questions which her work can answer will arise only in the future?[23] For example, here are the titles of the first five *AER* papers listed among my references: "Innovations in Large and Small Firms, An Empirical Analysis"; "The Welfare State and Competitiveness"; "Children and their Parents' Labor Supply, Evidence from Exogenous Variations in Family Size"; "Race and Gender Discrimination in Bargaining for a New Car"; "Momma's Got the Pill: How Anthony Compstock and Griswold vs. Connecticut Shaped U.S. Childbearing". How could their authors have determined the appropriate loss function?[24]

Admittedly, this response to Z-M is not entirely satisfactory because in doing the econometrics an econometrician, particularly an LSE econometrician, often has to make decisions based, at least in part, on significance tests, and those de-

22. It is not clear whether Z-M agree. In many of the instances they give where loss functions are needed one could allow the econometrician to function as the policymaker.

23. As Hoover and Siegler (2008a, 18) point out, one needs a loss function when deciding how strong to make a bridge, but not to set out the laws used to calculate its strength. Admittedly, postmodernists have criticized the claim that scientific statements can avoid all value judgments, and Rudner (1953) presents a criticism that zeros in on significance tests. However, even if one concedes that science cannot be purged completely of value judgments, one should do so to the extent one can. Even if there is no watertight dichotomy between value judgments and positive judgments, for practical purposes it is often a useful distinction because it is an instance of the division of labor. Policymakers are better at (and have more legitimacy in) making value judgments than econometricians are, and the econometrician's task here is merely to draw their attention to any significant value judgments implicit in the results he presents to them.

24. Moreover, evaluating policy options correctly requires more than combining econometric results with value judgments. It also requires judgments about the probable gap between the policy as proposed and as it is likely to emerge from the political and administrative caldrons, as well as its unintended effects on factors such as trust in government (see Colander 2001). A policymaker is probably better equipped to deal with such problems than is an econometrician.

cisions cannot be passed on to the policymaker. But since we do not have a relevant loss function for these cases, this qualification has no practical relevance.

The only workable solution is to designate the users of econometric estimates as the ones who should apply the appropriate loss function. Presumably Z-M's objection to this is that policymakers or other users may fail to apply the appropriate loss function and implicitly assume a symmetric one. This concern may well be justified. But the solution is to educate users of significance tests rather than to impose impossible demands on their providers.

Does a loss function have any relevance for deciding what to believe when it is just a matter of knowledge for knowledge's sake? (See Hoover and Siegler 2008a, 18.) The intuitively plausible answer in most cases is no, but how do we know that what on the surface seems like a belief that has no policy implications, will not ultimately have policy implications—perhaps implicitly, by changing the core of our belief set? But as a practical matter, in most cases where we just seek knowledge for knowledge's sake we do not know the loss function, so the symmetrical one implicitly assumed by significance tests is no more arbitrary than any other.

Presenting the results of significance tests

In economics and psychology the four most common ways of presenting the results of significance tests are t-values, *p*'s, F's, and confidence intervals. In response to Z-M's criticism of reporting just t-values, Hoover and Siegler (2008a) and Spanos (2008) point out that if readers are given, as they normally are, the point estimate and either the standard error or the t-value or else the confidence intervals, they can readily calculate the two other measures, so that it does not matter which one they are given. That is so. But it does not address the question of which measure is preferable, given that many readers are time constrained and therefore likely to look only at the measure explicitly provided.

Choosing between the measures

The choice between t-values and *p*'s and F's is inconsequential. What is important is the choice between any of them and confidence intervals. Confidence intervals have substantial advantages, and it is therefore not surprising that they have become more common in the medical literature, and that the American Psychological Association's Board of Scientific Affairs recommended that all estimates of size effects be accompanied by confidence intervals (see Fidler et al. 2004a; Stang, Poole, and Kuss 2010) First, confidence intervals make it much harder to ignore oomph since they are stated in terms of oomph (see McCloskey

and Ziliak (2008, 50); Hubbard and Armstrong (2006, 118)). Second, confidence intervals do not lend themselves as readily to wrong-way-round use as do t-values, F's, and *p*'s. While someone might mistakenly treat a hypothesis as disconfirmed because the t-value of the important regressor is, say, only 1.8, she is at least somewhat less likely to do so if told that its upper confidence interval shows it to be important. Confidence intervals thus reduce the hazards highlighted by Z-M's two main criticisms.

Third, when presenting t-values there is a temptation, often not resisted, to mine the data until $t \geq 2$ (see Brodeur et al. 2012). Presenting confidence intervals instead is likely to reduce this temptation. Fourth, the use of t-values or *p*'s generates an anomaly that confidence intervals are likely to avoid. It would be almost impossible to publish an applied econometrics paper that does not take account of sampling error. Yet, when an economist uses a coefficient generated by someone in a prior paper she usually employs only the point estimate, thus totally disregarding sampling error. If a paper presents confidence intervals, there is at least a chance that someone using its findings would undertake robustness tests using these confidence intervals.

In some econometric procedures confidence intervals are used frequently. Thus Hoover and Siegler (2008a, 20) point to their use in connection with impulse response functions. Hoover and Siegler note also that confidence intervals are the default setting for VARs in commonly used software packages and are typically reported when using autocorrelation or partial autocorrelation functions, as well as in connection with hazard functions and survivor functions. But in many other situations they are not reported. In my sample of *AER* papers many more papers provided t's than confidence intervals. One reason could be that for many papers confidence intervals would reveal the substantial imprecision of the paper's results, and thus their limited use for policymaking. Congress is not greatly helped if told that a stimulus somewhere between $100 billion and a $1 trillion is needed (cf. Johnson (1999, 769). And even papers that do not aim at direct policy conclusions or at forecasting can face a similar problem. To be told that the 1929 stock market decline accounted for somewhere between two and 80 percent of the subsequent decline in GDP would not satisfy one's curiosity.[25]

Is a standardized acceptance level appropriate?

The prevalence of a standardized acceptance level for *t*'s and *p*'s has the obvious advantage of eliminating the need for readers to make a decision, and it is also

25. Johnson (1999, 769) provides a whole list of invalid arguments that someone might give for not using confidence intervals.

easier for them to remember that a coefficient is significant than that it is, say, 2.3. But a standardized level also has two disadvantages. One is that an author, fearing that his papers will be rejected unless $t \geq 2$, has a strong incentive to ensure that it does so, even if it means running numerous and quite arbitrarily selected variants of his main regression—including ones that do not conform to the hypothesis as well as his original regression does—until eventually one yields a qualifying t-value.[26] Second, suppose that there are five independent studies of the effect of x on y. All find a positive effect, but their t-values are only 1.8, 1.7, 1.6, 1.5, and 1.4. If they are rejected because of this, the file-drawer problem in any subsequent meta-analysis is exacerbated, and a scientific result may thus be lost.[27] The availability of working-paper versions of articles that have been rejected because of low t-values does not eliminate this problem entirely because some researchers may not proceed to the working-paper stage if their results are not significant, or because the available meta-analyses may not cover working papers.

Conclusion

We should reject Z-M's claim that most of the significance testing done by economists is invalid, but we should also reject the idea that, at least in economics, all is well with significance tests. A basic problem with Z-M's claim is that, at least at times, they fail to realize that the purpose of a significance test is not just to test the maintained hypothesis, but to test whether the researcher's reliance on a sample instead of the entire universe invalidates her results, and that significance tests can therefore be treated as a requirement of data hygiene. Moreover, Spanos (2008) may well be right that statistical misspecification is a more serious problem for econometrics than is the misuse of significance tests. The same may perhaps be true for computational errors (see Dewald, Thursby, and Anderson 1986; McCullough and Vinod 1999) and even for inattention to the meaning and accuracy of the data (see Cargill 2012; Corrado, Hulton, and Sichel 2006; Reinsdorf 2004). But misuse of significance tests is an easier problem to cure.

Nonetheless, Z-M are right in saying that one must guard against substituting statistical for substantive significance and that economists should pay more attention to substantive significance. They are also right in criticizing the wrong-way-round use of significance tests. In the testing of maintained hypotheses this error

26. Keuzenkamp and Magnus (1995, 18) report that the *Journal of the Royal Statistical Society* (JRSS) has been called the *Journal of Statistically Significant Results*.

27. For estimates of how seriously the results of meta-analytic studies are distorted by the unwillingness of authors to submit papers with low t values, and the unwillingness of journals to publish such papers, see Sterling, Rosenbaum, and Weinkam (1995) and Brodeur et al. (2012) and the literature cited therein.

is both severe enough and occurs frequently enough to present a serious—and inexcusable—problem. (Perhaps editors should specifically ask referees to check the use of significance tests.) And the error is probably much more widespread when deciding whether to adjust for serial correlation, heteroscedasticity, etc. Someone who wants to argue that the great majority of time-series econometric papers are flawed due to a misuse of significance tests should focus on congruity adjustments. Z-M are also right that an explicit loss function needs to be used when making policy decisions. But the econometrician is generally not able to do so and should leave that up to the policymaker. Also, because many readers are harried, Z-M are correct in their claim that confidence intervals are usually more informative than are t-values or p's.

Moreover, in countering the mechanical way in which significance tests are often used, and in introducing economists to the significance-test literature in other fields, Z-M have rendered valuable services. But there is a danger that their reliance on some flawed arguments, as well as their vehemence and overreach, will tempt economists to dismiss their work.

Appendix A: Critique of Z-M's criteria for evaluating significance tests

Note: All quotations are from Ziliak and McCloskey (2008b). In the original the first sentence of each paragraph is in italics. After quoting and sometimes elaborating the criterion, I give my evaluation.

Criterion 1: "Does the article depend on a small number of observations such that statistically 'significant' differences are *not* forced by the large number of observations?" (p. 67) *Evaluation*: This is not an appropriate criterion for capturing a misuse of significance tests. The question that these tests are intended to answer is whether we can reject the possibility that the observed difference can reasonably be attributed merely to sampling error. If it cannot, it does not matter whether that is due to the difference (or the regression coefficient) being large relative to the variance, or to the sample being large. Z-M justify their criterion by saying that we know that with a large enough sample every difference will be significant. But even if that were true, it would be irrelevant, because the function of significance tests is to tell us whether we can claim that the existence of a difference (or the nonzero value of a coefficient) in our sample implies the same for the universe, regardless of whether it does so because the sample size is large, or the difference is large relative to the variance.

Criterion 2: "Are the units and the descriptive statistics for all regression variables included? … No one can exercise judgment as to whether something is importantly large or small when it is reported without units or a scale along which to judge them large or small…" (67) *Evaluation*: What matters is not the comprehensibility of "all" variables, but only of the strategic ones. Hence, this criterion is too tough. Second, authors need not specify the units and descriptive statistics if they are obvious. Apart from that, Z-M have a valid point, but one that has no discernible relation to significance tests per se. If I publish a table for which the units are neither defined nor obvious (say a table in which the units are $10) I am failing to inform readers about my findings, whether I run significance tests or not.

Criterion 3: "Are the coefficients reported in elasticity form, or in some interpretable form relevant for the problem at hand, so that the reader can discern the economic impact? … [O]ften an article will not give the actual magnitude of the elasticity but merely state with satisfaction its statistical significance." (67) *Evaluation*: This is often a valid criterion when applied not to every regression coefficient that is presented but only to the strategic ones. But even then, not always. As discussed in

the text, there are cases in which it is the t-value and not the oomph that matters. So this criterion is valid only some of the time.

Criterion 4: "Are the proper null hypotheses specified? Sometimes the economists will test a null of zero when the economic question entails a null quite different from zero…." (68) Z-M's example is testing whether the income elasticity of money is unity. *Evaluation*: This criterion is valid only if the article, after mentioning the (irrelevant) result of testing against zero, does not go on to perform the proper test as well.

Criterion 5: "Are the coefficients carefully interpreted?" (68) Z-M's example is a regression of a person's weight on his height and miles walked per week, where the height variable is statistically significant and the miles-walked variable is not, though its coefficient is large. These results do not imply that if you want to lose weight without exercising just grow taller. *Evaluation*: Yes, this is right, but it is a problem of whether the regression results have been interpreted correctly, and not of significance tests per se. The mistake would be there even if the author had never run a significance test and relied entirely on the large oomph of height.

Criterion 6: "Does the article refrain from reporting *t*- or *F*- statistics or standard errors even when a test of significance is not relevant? A No on this question is another sign of canned regression packages taking over the mind of the scientist." (68-69) Z-M give the example of applying a significance test when the sample consists of the entire universe, a problem I discuss in footnote 5. *Evaluation*: Yes, reporting meaningless measures should be avoided.

Criterion 7: "Is statistical significance at its first use merely one of multiple criteria of 'importance' in sight? Often the first use will be at the crescendo of the article, the place where the author appears to think she is making the crucial factual argument. But statistical significance does not imply substantive significance. … Articles were coded Yes if statistical significance played a second or lower-order role, at any rate below the primary considerations of substantive significance." (69) *Evaluation*: As discussed in the text, there are cases in which statistical significance *should* play a primary role. Second, why is it necessarily wrong to stress statistical significance at the first use or crescendo if the article adequately discusses substantive significance at another point? An author may well want to discuss first whether to believe that the observation, say a positive regression coefficient in her sample, reliably tells us anything about the universe, or could just be dismissed as perhaps due to sampling error, and discuss substantive significance later.

Criterion 8: "Does the article mention the power of the test?" (69) *Evaluation*: If the test rejects the hypothesis there is no reason why its power need be mentioned. Hence it is not relevant in some of the cases.

Criterion 9: "If the article mentions power, does it do anything about it?" (69) *Evaluation*: This criterion partially overlaps with the previous one, and is subject to the same criticism. Treating them as separate criteria gives this criterion sometimes a double weight.

Criterion 10: "Does the article refrain from 'asterisk econometrics,' that is, ranking the coefficients according to the absolute size of their *t*-statistics?" (70) *Evaluation*: Presumably what Z-M mean with "ranking" is the order in which the variables and their coefficients are listed in a table. If so, while such a ranking may enhance a reader's inclination to overvalue statistical significance, it does not itself amount to an incorrect use of significance tests, and is more a matter of style.

Criterion 11: "Does the article refrain from 'sign econometrics,' that is, noting the sign but not the size of coefficients? The distribution-free 'sign test' for matched pairs is on occasion scientifically meaningful. Ordinarily sign alone is not *economically* significant, however, unless the magnitude attached to the sign is large or small enough to matter." (70, italics in original) *Evaluation*: As shown in the text, there is more scope for sign tests in economics than Z-M allow. However, in other cases this is a valid criterion. But it is unlikely that there are many such cases, because it would be strange if the table giving the sign does not also give the coefficient.

Criterion 12: "Does the article discuss the size of the coefficients at all? Once regression results are presented, does the article ask about the *economic* significance of the results?" (70, italics in original) *Evaluation*: As just mentioned there is some scope for sign-only significance tests. However, for many (probably most) papers, Z-M are right; the size of coefficients does matter, and it would often help the reader if it were discussed. But does it *have* to be discussed? It is not clear how much convenience to the reader an author is obligated to provide. If the economic meaning of the coefficient is complex, then an efficient division of labor requires the author to discuss it. However, in some (many?) cases economic significance may be too obvious to require discussion, e.g. an elasticity of hours worked with respect to the wage rate of 0.001. Or the purpose of the article may be to reject previously published papers that do discuss the economic significance of their coefficients, which therefore does not have to be discussed again. Thus it is not clear whether an article should be faulted, and by how much, for presenting oomph only in tables.

Criterion 13: "Does the article discuss the scientific conversation within which a coefficient would be judged 'large' or 'small'?" (71) *Evaluation*: This is not always needed. It may be obvious, or there may not be much of a prior conversation. Moreover, as explained in the text, in some cases only the sign or t-values matter.

Criterion 14: "Does the article refrain from choosing variables for inclusion in its equations solely on the basis of statistical 'significance'? … [T]here is no scientific reason—unless a reason is provided, and it seldom is—to drop an 'insignificant' variable. If the variable is important substantively but is dropped from the regression because it is Fisher-insignificant, the resulting fitted equation will be misspecified…." (71) *Evaluation*: This is discussed in the text.

Criterion 15: "Later, after the crescendo, does the article refrain from using statistical significance as the criterion of scientific importance? Sometimes the referees will have insisted unthinkingly on a significance test, and the appropriate *t*'s and *F*'s have therefore been inserted." (72) *Evaluation*: Without asking the authors, we cannot know whether the inclusion of a significance test after the crescendo was the author's own idea, or was forced on her. But why does the origin of the significance test matter? If it shows that the estimated substantive significance of a coefficient is not just the product of sampling error, it is useful regardless of what prompted it.

Criterion 16: "Is statistical significance portrayed as decisive, a conversation stopper, conveying a sense of an ending?" (72) *Evaluation*: Z-M treat a positive answer as an error. Once again, in some situations it is not. Suppose someone published a paper relating the growth rates of countries to the level of their corporate income tax, and found a negative correlation. If you now write a paper showing that the difference in the growth rates is not statistically significant, and should therefore not be treated as convincing evidence on the appropriate level of corporate income taxes, are you making a mistake?

Criterion 17: "Does the article ever use an independent simulation—as against a use of the regression coefficients as inputs into further calculations—to determine whether the coefficients are reasonable?" (72) *Evaluation*: Such simulations may be useful in some perhaps many cases, but should every failure to use simulations count as a fault? As Wooldridge (2004) points out, useful simulations are sometimes not feasible.

Criterion 18: "In the concluding sections is statistical significance separated from policy, economic, or scientific significance? In medicine and epidemiology and especially psychology the concluding sections are often sizeless summaries of significance tests reported earlier in the article. Significance this, significant that. In

economics, too." (72-73) *Evaluation*: I doubt that in economics this is an accurate description of the concluding sections of many articles. And in those cases where significance is the point at issue, it should not count against the article.

Criterion 19: "Does the article use the word *significant* unambiguously?" (73) *Evaluation*: Yes, ambiguity is bad. But that does not mean that the article misuses significance tests.

In summary: Although any attempt to fit the results of this evaluation of Z-M's criteria into a few broad classes requires some judgment calls, I would classify seven of the nineteen criteria (1, 2, 5, 7, 10, 15 and 19) as invalid, ten (3, 4, 8, 9, 11, 12, 13, 16, 17 and 18) as valid in some cases, but not in others, one (14) as debatable, and another (6) as only weakly relevant because little damage results from not satisfying it. However, this judgment is the product of looking at each criterion in isolation and of applying it in a fairly mechanical way to all significance tests. A more nuanced procedure that allows for the fact that not all nineteen criteria are applicable to every significance test, and that different criteria have different weights in particular cases, might result in a much more favorable judgment. But that is similar to looking at the Gestalt of the significance test, as is done in the text.

Appendix B: Reappraising eleven papers that Z-M rank "poor" or "very poor" with respect to their use of significance tests

Ziliak and McCloskey (2008b, 91-92) classify papers published in the *AER* during the 1990s into five categories with respect to their use of significance tests: "exemplary" (6 percent); "good" (14%); "fair" (22%); "poor" (37%); and "very poor" (20%). Eleven of the papers that they rank "poor" or "very poor" are also in my sample, and I discuss them here, classifying them into four categories: "good", "fair", "marginal", and "bad". I start with the ones that Z-M rank lowest, so that the first three are ones that Z-M classify as "very poor", and the others are papers they classify as "poor".

1. S. Lael Brainard (1997), "An Empirical Assessment of the Proximity-Concentration Trade-Off between Multinational Sales and Trade"

Brainard evaluates the proximity-concentration hypothesis that predicts that firms expand across national borders when the benefits of closer access to their customers exceed the benefits obtainable from economies of scale. He builds a model embodying this hypothesis and then runs the required regressions. In the text he only discussed the signs and significance of the variables, but his tables provide the coefficients. And since his regressions are in logs, these coefficients are easy to interpret. All the same, since hasty readers may not bother to look at these tables, it would have been better to discuss the oomph of the strategic variables in the text. But Z-M's grade of "very poor" seems unjustified, and a grade of "good", or at the least "fair", seems more appropriate.

2. Stephen Trejo (1991), "The Effect of Overtime Pay Regulation on Worker Compensation"

To see whether regulations governing overtime pay, such as the time-and-a-half rule, affect total labor compensation, or whether firms offset the requirement to pay more for overtime by lowering regular wage rates, Trejo first compares the extent to which firms comply with overtime-pay regulations for workers at and above the minimum wage since firms are much more likely to lower regular wage rates that are above the minimum wage than those that are at minimum-wage level. Trejo therefore compares compliance rates with the overtime-pay rule for workers at and above the minimum wage. He finds a statistically significant difference, with firms complying less frequently with the time-and-a-half requirement when regular wages are at the minimum level. This is consistent with a model in which

firms—when minimum wage laws do not prevent it—cut regular wages to compensate for having to pay overtime rates. Trejo found that "the estimated effects of this variable are relatively large… Within the covered sector, minimum-wage workers are associated with 3–9 percentage points lower probability of being paid an overtime premium" (729). Moreover, the harder it is for firms to reduce regular wage rates to offset paying time and a half for overtime, the greater is their incentive not to exceed the forty-hour limit. One should therefore find greater bunching of workers at the forty-hour level for firms that pay just the minimum wage than for firms that have more scope to reduce regular wages. And Trejo's regressions confirm that such bunching occurs, with the coefficient of the relevant variable being both "positive and relatively large" (731). He also investigates whether straight-time pay adjusts fully to offset the requirement for overtime pay. It does not. But still, the coefficient that shows the adjustment of straight-time pay is "negative and statistically significant" (735). He leaves it to the reader to obtain its magnitude from the accompanying tables. In a final set of regressions using weekly data, Trejo finds that the coefficients of a variable whose (positive) significance and importance would contradict his hypothesis, do not "achieve statistical significance, are negative in 1974, and in the other years are always less than a third of the value predicted by" the rival theory (737). All in all, the treatment of significance tests in this paper deserves a grade of "good".

3. Randall Kroszner and Raghuram Rajan (1994), "Is the Glass-Steagall Act Justified? A Study of U.S. Experience with Universal Banking"

The Glass-Steagall Act of 1933 prohibited commercial banks from underwriting and trading in corporate securities. A major reason was to avoid potential conflicts of interest, such as a bank taking advantage of its greater information about its borrowers by underwriting securities issued by its weaker borrowers, so that these borrowers can repay their loans to the bank. Kroszner and Rajan investigate whether banks actually succeeded in taking such advantage of inside information by seeing whether securities underwritten by commercial banks or their affiliates performed worse than those issued by investment banks. To do that they constructed 121 matched pairs of security issues underwritten by commercial banks and by investment banks. What they found strongly contradicts the asymmetric-information hypothesis; securities issued by investment banks suffered about 40 percent more frequent defaults than those issued by commercial banks and their affiliates. When measured by the dollar volume of defaults, the difference in favor of commercial banks and their affiliates is even greater. Kroszner and Rajan also show that the difference in default rates is even greater for bonds below investment grade, which is inconsistent with the hypothesis that commercial banks were taking advantage of naïve investors. For some of their

regressions they provide both the significance and oomph in their text, while for some others they provide the oomph only in their tables. This is justified because their aim was to challenge the then widely accepted hypothesis that allowing commercial banks and their affiliates to underwrite securities creates a conflict of interest. For that, it suffices to show that the relevant differences have the wrong sign. This paper therefore deserves at least a "good".

4. Robert Feenstra (1994), "New Product Varieties and the Measurement of International Prices"

This is primarily a theoretical paper on how to incorporate new product varieties into demand functions for imports, but it illustrates the procedure by estimating the income elasticity for six U. S. imports. Feenstra cites these elasticity estimates, and thus the oomph, extensively in the text, not just in the tables. He does, however, at one point use a wrong-way-round significance test. But this point is not important for the paper, and I therefore classify it as "marginal".

5. Jeffrey Fuhrer and George Moore (1995), "Monetary Policy Trade-Offs and the Correlation Between Nominal Interest Rates and Real Output"

This paper estimates a small model that explains the observed relation between changes in the short-term interest rate and output, using both VARs and a structural model. It presents its results mainly by charts of autocorrelation functions and autocovariance functions, so that no t-values are mentioned and a reference to "significance" appears only once. That is when Fuhrer and Moore report that they chose the lag lengths for the regressors by reducing "the lag length until the last lag remains statistically significant and the residuals appear to be uncorrelated" (220). Since, as discussed above, that is a questionable use of significance tests, I put this paper into the "marginal" bin, though Fuhrer and Moore should not be castigated for using what is a standard procedure.

6. Ian Ayers and Peter Siegelman (1995), "Race and Gender Discrimination in Bargaining for a New Car"

Ayers and Siegelman sent black and white, and male and female testers to Chicago area car dealers, and compared the prices they were offered. Their regressions show that the race and gender of testers "strongly influence both the initial and final offers made by sellers" (309). Throughout the text they cite numerous oomphs, for example: "For black males, the final markup was 8–9 percentage points higher than for white males; the equivalent figures are 3.5–4 percentage points for black females and about 2 percentage points for white females" (313). Moreover, Ayres and Siegelman sometimes cite oomph even when the t-statistics are far from significant. But at one, minor point, they do use significance tests the wrong way round. Hence, I give their paper a "marginal" grade.

7. Edward Wolff (1991), "Capital Formation and Productivity Convergence Over the Long Run"

Wolff investigates the international convergence of productivity and tests three explanatory hypotheses: the catch-up hypothesis, which implies that the further a country lags technologically, the faster will be its rate of catch-up, an alternative hypothesis that convergence in labor productivities is due to convergence in factor intensities, and a third hypothesis that there exist positive effects of capital accumulation on technological progress. Wolff runs several regressions. While providing the magnitude of regression coefficients in his tables, in his text Wolff discusses primarily their signs and significance. And at one rather peripheral point he excludes two potential regressors because their coefficients are not significant, even though his sample is fairly small. Hence, one might argue that he uses these significance tests the wrong way round. But, given the need to limit somehow the huge number of regressors that one might potentially include (and *perhaps* also the frequency with which insignificant variables are dropped in economics), it seems that "marginal" is a more appropriate grade than the "poor" that Z-M label it.

8. Kenneth Hendricks and Robert Porter (1996), "The Timing and Incidence of Exploratory Drilling on Offshore Wildcat Tracts"

When wildcat drillers bid successfully on drilling leases, they have to decide whether to incur the cost of actually drilling on these leases. In making this decision they look at what owners of other leases in the area are doing, since the productivity of wells within the same area is likely to be correlated. Each leaseholder therefore has an incentive to wait and see how successful others are. How is this problem resolved in practice? After developing the required theory, Hendricks and Porter present tobit regressions of the logs of the discounted annual revenue from drilling tracts. They provide the regression coefficients and t-values in their tables, while in their text they take up the important t-values and some of the regression coefficients. That they discuss only some of the coefficients in the text is not a serious problem because the coefficients as given in the tables are easy to interpret since their variables are measured in logs and have straightforward meanings. Hence, this paper deserves a "good".

9. Albert Alesina and Robert Perotti (1997), "The Welfare State and Competitiveness"

The basic idea of this paper is that a rise in taxes on labor to finance enhanced benefits for pensioners or the unemployed causes unions to press for higher wages, which results in a loss of competitiveness. The distortions thus introduced are greater the stronger are unions, until we reach the point when wage negotiations move to the national level where unions internalize the negative effects of their policies. After developing a model built on these insights, Alesina and Perotti

estimate it for a panel of the manufacturing sectors of 14 OECD countries. In doing so they discuss extensively, not just the t-values, but also the regression coefficients. Since these coefficients have clear-cut meanings, the reader is well informed about oomph. And since there are no instances of wrong-way-round significance tests, the paper deserves a "good", not the "poor" that Z-M give it.

10. Jordi Gali (1999), "Technology, Employment and the Business Cycle"
Gali presents a test of real business cycle theory, focusing on the theory's (counterfactual) positive correlation between labor productivity and hours worked, a correlation that can potentially be explained by other shocks. He builds a VAR model embodying both types of shocks. In this model a technological shock must affect labor productivity *permanently*, and Gali uses that as his identifying restriction. In presenting his results he not only gives the regression coefficients in his tables, and presents numerous impulse response functions, but also frequently discusses oomph in his text. There is, however, one place where he uses significance tests wrong way round. This is in deciding whether to adjust for cointegration. When dealing with U.S. data (his main results) he correctly runs his regressions in both ways (and gets similar results), but does not do that when dealing with foreign data. All in all, his use of significance tests deserves a "fair".

11. Robert Mendelsohn, William Nordhaus, and Daigee Shaw (1994), "The Impact of Global Warming on Agriculture: A Ricardian Analysis"
The usual way economists have studied the impact of climate change is to fit production functions containing climate variables for various crops. But as Mendelsohn, Nordhaus, and Shaw point out, such a technological approach overestimates the losses from climate change, because it ignores that farmers can respond by changing both their production technology and their crop mix. Instead, the authors allow for adaptations, such as a shift to entirely new uses for land, by adopting a "Ricardian" approach that looks at how differences in climate affect, not the output of particular crops, but the rent or value of farmland. To do that they regress average land value and farm revenue for all counties in the lower 48 states on climate and non-climate variables. Since in presenting their results they put much greater stress on oomph (that is, on changes in the dollar value of harvests as climate changes) than on t-values, and since they do not use significance tests wrong-way-round, this paper should be graded "good".

Thus in my alternative classification of these 11 papers, six receive a "good", one a "fair", and four a "marginal". It is highly likely that another economist would come up with different grades for some of the papers (and probably so would I if I would repeat the exercise), but he is most unlikely to come up with grades that are anywhere near Z-M's harsh results.

References

Acs, Zoltan, and David Audretsch. 1988. Innovation in Large and Small Firms: An Empirical Analysis. *American Economic Review* 78(Sept.): 678-690.

Alesina, Alberto, and Roberto Perotti. 1997. The Welfare State and Competitiveness. *American Economic Review* 87(Dec.): 921-939.

Altman, Douglas. 1980. Statistics and Ethics in Medical Research. *British Medical Journal* 281(Nov.): 1336-1338.

Altman, Morris. 2004. Introduction. *Journal of Socio-Economics* 33(Nov.): 523-525.

Angrist, Joshua, and William Evans. 1998. Children and their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size. *American Economic Review* 88(June): 450-477.

Anonymous. 2012. Report on "Ziliak and McCloskey's Criticism of Significance Tests: A Damage Assessment". Referee report on this paper.

Artuç, Erhan, Shubham Chaudhuri, and John McLaren. 2010. Trade Shocks and Labor Adjustment: A Structural Empirical Approach. *American Economic Review* 100(3): 1008-1045.

Ayers, Ian, and Peter Siegelman. 1995. Race and Gender Discrimination in Bargaining for a New Car. *American Economic Review* 85(June): 304-321.

Bailey, Martha. 2010. Momma's Got the Pill: How Anthony Compstock and Griswold v. Connecticut Shaped U. S. Childbearing. *American Economic Review* 100(Mar.): 98-129.

Bardhan, Pranab, and Dilip Mookherjee. 2010. Determinants of Redistributive Policies: An Empirical Analysis of Land Reforms in West Bengal, India. *American Economic Review* 100(Sept.): 1572-1600.

Berg, Nathan. 2004. No Decision Classification: An Alternative to Testing for Statistical Significance. *Journal of Socio-Economics* 33(Nov.): 631-650.

Blanchard, Olivier. 1989. Traditional Interpretations of Macroeconomic Fluctuations. *American Economic Review* 79(Dec.): 1146-1164.

Blaug, Mark. 1980. *The Methodology of Economics*. New York: Cambridge University Press.

Bloom, David, and Christopher Cavanagh. 1986. An Analysis of the Selection of Arbitrators. *American Economic Review* 76(June): 408-422.

Borjas, George. 1987. Self-Selection and the Earnings of Immigrants. *American Economic Review* 77(Sept.): 531-553.

Borjas, George. 1995. Ethnicity, Neighborhoods and Human Capital Externalities. *American Economic Review* 85(June): 365-390.

Brainard, S. Lael. 1997. An Empirical Assessment of the Proximity-Concentration Trade-Off Between Multinational Sales and Trade. *American Economic Review* 87(Sept.): 520-544.

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg. 2012. Star Wars: The Empirics Strike Back. *Paris School of Economics Working Paper* No. 2012-29. Paris-Jourdan Sciences Economiques (Paris). **Link**

Cargill, Thomas. 2012. A Critical Assessment of Measures of Central Bank Independence. *Economic Inquiry*, forthcoming. **Link**

Carmichael, Jeffrey, and Peter Stebbing. 1983. Fisher's Paradox and the Theory of Interest. *American Economic Review* 73(Sept.): 619-630.

Chandra, Amitabh, Jonathan Gruber and Robin McKnight. 2010. Patient Cost Sharing and Hospitalization Offsets in the Elderly. *American Economic Review* 100(Mar.): 193-213.

Chen, Yan, F. Maxwell Harper, Joseph Konstan, and Sherry Xin Li. 2010. Social Comparisons and Contributions to Online Communities: A Field Experiment on MovieLens. *American Economic Review* 100(4): 1358-1398.

Cohen, Jacob. 1994. The Earth Is Round ($p < 0.05$). *American Psychologist* 49(Dec.): 997-1003.

Colander, David. 2001. *The Lost Art of Economics*. Cheltenham, UK: Edward Elgar.

Colander, David. 2011. Creating Humble Economists: A Code of Ethics for Economists. *Middlebury Economics Working Paper* 11-03. Department of Economics, Middlebury College (Middlebury, Vt.). **Link**

Conley, Timothy, and Christopher Udry. 2010. Learning About a New Technology: Pineapples in Ghana. *American Economic Review* 100(Mar.): 35-69.

Corrado, Carol A., Charles R. Hulten, and Daniel E. Sichel. 2006. Intangible Capital and Economic Growth. *NBER Working Paper* No. 11948. National Bureau of Economic Research (Cambridge, Mass.). **Link**

Dafny, Leemore. 2010. Are Health Insurance Markets Competitive? *American Economic Review* 100(Sept.): 1399-1431.

Darby, Michael. 1982. The Price of Oil and World Inflation and Recession. *American Economic Review* 72(Sept.): 738-751.

Dewald, William, Jerry Thursby, and Richard Anderson. 1986. Replication in Economics: The Journal of Money, Credit and Banking Project. *American Economic Review* 76(Sept.): 587-603.

Ellison, Glenn, Edward Glaeser, and William Kerr. 2010. What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns. *American Economic Review* 100(June): 1195-1214.

Elliott, Graham, and Clive Granger. 2004. Evaluating Significance: Comments on "Size Matters." *Journal of Socio-Economics* 33(Nov.): 547-550.

Engsted, Tom. 2009. Statistical vs. Economic Significance in Economics and Econometrics: Further Comments on McCloskey and Ziliak. *Journal of Economic Methodology* 16(Dec.): 393-408.

Evans, David, and James Heckman. 1984. A Test for Subaddivity of the Cost Function with an Application to the Bell System. *American Economic Review* 74(Sept.): 615-623.

Feenstra, Robert. 1994. New Product Varieties and the Measurement of International Prices. *American Economic Review* 84(Mar.): 157-177.

Fidler, Fiona, Neil Thomason, Geoff Cumming, Sue Finch, and Joanna Leeman. 2004a. Editors Can Lead Researchers to Confidence Intervals, but Can't Make Them Think: Statistical Reform Lessons from Medicine. *Psychological Science* 15(Feb.): 119-126.

Fidler, Fiona, Geoff Cumming, Mark Burgman, and Neil Thomason. 2004b. Statistical Reform in Medicine, Psychology and Ecology. *Journal of Socio-Economics* 33(Nov.): 615-630.

Forsythe, Robert, Forrest D. Nelson, George R. Neumann, and Jack Wright. 1992. Anatomy of an Experimental Political Stock Market. *American Economic Review* 82(5): 1142-1161.

Fowlie, Meredith. 2010. Emissions Trading, Electricity Restructuring, and Investment in Pollution Abatement. *American Economic Review* 100(June): 837-869.

Friedman, Milton. 1957. *A Theory of the Consumption Function*. New York: Columbia University Press.

Froyen, Richard, and Roger Waud. 1980. Further International Evidence on the Output-inflation Tradeoffs. *American Economic Review* 70(Mar.): 409-421.

Fuhrer, Jeffrey, and George Moore. 1995. Monetary Policy Trade-offs and the Correlation Between Nominal Interest Rates and Real Output. *American Economic Review* 85(Mar.): 219-239.

Gali, Jordi. 1999. Technology, Employment and the Business Cycle: Do Technology Shocks Explain Aggregate Fluctuations? *American Economic Review* 89(Mar.): 249-271.

Garber, Peter. 1986. Nominal Contracts in a Bimetallic Standard. *American Economic Review* 76(Dec.): 1012-1030.

Gibbard, Allan, and Hal Varian. 1978. Economic Models. *Journal of Philosophy* 75(Nov.): 665-667.

Gigerenzer, Gerd. 2004. Mindless Statistics. *Journal of Socio-Economics* 33(Nov.): 587-606.

Gill, Jeff. 1999. The Insignificance of Null Hypothesis Significance Testing. *Political Research Quarterly* 52(Sept.): 647-674.

Ham, John, Jan Svejnar, and Katherine Terrell. 1998. Unemployment and the Social Safety Net During Transitions to a Market Economy: Evidence from the Czech and Slovak Republics. *American Economic Review* 88(Dec.): 1117-1142.

Haller, Heiko, and Stefan Krauss. 2002. Misinterpretations of Significance: A Problem Students Share with Their Teachers? *Methods of Psychological Research Online* 7(1).

Harlow, Lisa, Stanley Mulaik, and James Steiger. 1997. *What If There Were No Significance Tests?* Mahwah, N.J.: Lawrence Erlbaum Associates.

Harrison, Ann, and Jason Scorse. 2010. Multinationals and Anti-Sweatshop Activism. *American Economic Review* 100(Mar.): 247-273.

Hendricks, Kenneth, and Robert Porter. 1996. The Timing and Incidence of Exploratory Drilling on Offshore Wildcat Tracts. *American Economic Review* 86(June): 388-407.

Hoover, Kevin. 2011. The Role of Hypothesis Testing in the Molding of Econometric Models. Working paper.

Hoover, Kevin, and Mark Siegler. 2008a. Sound and Fury: McCloskey and Significance Testing in Economics. *Journal of Economic Methodology* 15(Mar.): 1-38.

Hoover, Kevin, and Mark Siegler. 2008b. The Rhetoric of "Signifying Nothing": A Rejoinder to Ziliak and McCloskey. *Journal of Economic Methodology* 15(Mar.): 57-68.

Hoover, Kevin, and Steven Sheffrin. 1992. Causality, Spending and Taxes: Sand in the Sandbox or Tax Collector for the Welfare State? *American Economic Review* 82(Mar.): 225-248.

Horowitz, Joel. 2004. Comments on "Size Matters." Journal of Socio-Economics 33(Nov.): 551-554.

Hubbard, Raymond, and S. Scott Armstrong. 2006. Why We Really Don't Know What Statistical Significance Means: Implications for Educators. *Journal of Marketing Education* 28(Aug.): 114-120.

Johnson, William, and Jonathan Skinner. 1986. Labor Supply and Marital Separation. *American Economic Review* 76(June): 455-469.

Johnson, Douglas. 1999. The Insignificance of Statistical Significance Testing. *Journal of Wildlife Management* 63(July): 763-772.

Joskow, Paul. 1987. Contract Duration and Relationship-Specific Investments: Empirical Evidence from Coal Markets. *American Economic Review* 77(Mar.): 168-185.

Keuzenkamp, Hugo, and Jan Magnus. 1995. On Tests and Significance in Econometrics. *Journal of Econometrics* 67(1): 5-24.

Krämer, Walter. 2011. The Cult of Statistical Significance: What Economists Should and Should Not Do to Make Their Data Talk. *RatSWD Working Papers* 176. German Data Forum (Berlin). **Link**

Kroszner, Randall, and Raghuram Rajan. 1994. Is the Glass-Steagall Act Justified? A Study of the U.S. Experience with Universal Banking Before 1933. *American Economic Review* 84(Sept.): 810-832.

LaLonde, Robert. 1986. Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *American Economic Review* 76(Sept.): 604-620.

Landry, Craig E., Andreas Lange, John A. List, Michael K. Price, and Nicholas G. Rupp. 2010. Is a Donor in Hand Better than Two in the Bush? Evidence from a Natural Field Experiment. *American Economic Review* 100(3): 958-983.

Leamer, Edward. 2004. Are the Roads Red? Comments on "Size Matters." *Journal of Socio-Economics* 33(Nov.): 555-557.

Lerner, Josh, and Ulrike Malmendier. 2010. Contractibility and the Design of Research Agreements. *American Economic Review* 100(Mar.): 214-246.

Leth-Petersen, Søren. 2010. Intertemporal Consumption and Credit Constraints: Does Total Expenditure Respond to an Exogenous Shock to Credit? *American Economic Review* 100(June): 1080-1103.

Mayer, Thomas. 1972. Permanent Income, Wealth, and Consumption. Berkeley: University of California Press.

Mayer, Thomas. 1980. Economics as an Exact Science: Realistic Goal or Wishful Thinking? *Economic Inquiry* 18(Apr.): 165-178.

Mayer, Thomas. 1993. Truth Versus Precision in Economics. Aldershot, UK: Edward Elgar.

Mayer, Thomas. 2001. Misinterpreting A Failure to Disconfirm as a Confirmation. *University of California, Davis, Department of Economics Working Papers* 01-08. University of California, Davis. **Link**

Mayo, Deborah. 1996. *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.

Mayo, Deborah, and Aris Spanos. 2006. Severe Testing as a Basic Concept in a Neyman-Pearson's Philosophy of Induction. *British Journal for the Philosophy of Science* 57(2): 323-357.

McCloskey, D. N. 1985. *The Rhetoric of Economics*. Madison: University of Wisconsin Press.

McCloskey, D. N. 2008. Private communication.

McCloskey D. N., and Stephen Ziliak. 2008. Signifying Nothing: Reply to Hoover and Siegler. *Journal of Economic Methodology* 15(Mar.): 39-56.

McCullough, D. B., and H. D. Vinod. 1999. The Numerical Reliability of Econometric Software. *Journal of Economic Literature* 37(June): 633-665.

Mendelsohn, Robert, William Nordhaus, and Daigee Shaw. 1994. The Impact of Global Warming on Agriculture: A Ricardian Analysis. *American Economic Review* 84(Sept.): 753-771.

Mian, Atif, Amir Sufi, and Francesco Trebbi. 2010. The Political Economy of the U.S. Mortgage Default Crisis. *American Economic Review* 100(Dec.): 1967-1998.

Mishkin, Frederic. 1982. Does Anticipated Aggregate Demand Policy Matter: Further Econometric Results. *American Economic Review* 72(Sept.): 788-802.

Morrison, Denton, and Ramon Hankel. 1970. *The Significance Test Controversy: A Reader*. Chicago: Aldine.

Nickerson, Raymond. 2000. Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy. *Psychological Methods* 5(2): 241-301.

O'Brien, Anthony. 2004. Why Is the Standard Error of Regressions So Low Using Historical Data? *Journal of Socio-Economics* 33(Nov.): 565-570.

Pashigian, Peter. 1988. Demand Uncertainty and Sales: A Study of Sales and Markdown Pricing. *American Economic Review* 78(Dec.): 936-953.

Pontiff, Jeffrey. 1997. Excess Volatility and Closed- End Funds. *American Economic Review* 87(Mar.): 155-169.

Pollard, P., and J. T. E. Richardson. 1987. On the Probability of Making Type I Errors. *Psychological Bulletin* 102(1): 159-163.

Reinsdorf, Marshall. 2004. Alternative Measures of Personal Saving. *Survey of Current Business* 84(Sept.): 17-27.

Robinson, Daniel, and Howard Wainer. 2002. On the Past and Future of Null Hypothesis Significance Testing. *Journal of Wildlife Management* 66(2): 263-271.

Romer, Christina. 1986. Is Stabilization of the Postwar Economy a Figment of the Data? Estimates Based on a New Measure of Fiscal Shocks. *American Economic Review* 76(June): 314-334.

Romer, Christina, and David Romer. 2010. The Macroeconomic Effects of Tax Changes: Estimates Based on a New Measure of Fiscal Shock. *American Economic Review* 100(June): 763-801.

Rudner, Richard. 1953. The Scientist Qua Scientist Makes Value Judgments. *Philosophy of Science* 20(Jan.): 1-6.

Sachs, Jeffrey. 1980. The Changing Cyclical Behavior of Wages and Prices, 1890-1976. *American Economic Review* 70(Mar.): 78-90.

Sauer, Raymond, and Keith Leffler. 1990. Did the Federal Trade Commission's Advertising Substantiation Program Promote More Credible Advertising? *American Economic Review* 80(Mar.): 191-203.

Spanos, Aris. 2008. Review of Stephen T. Ziliak and Deirdre N. McCloskey's *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice and Lives*. *Erasmus Journal for Philosophy and Economics* 1(Autumn): 154-164. **Link**

Stang, Andreas, Charles Poole, and Oliver Kuss. 2010. The Ongoing Tyranny of Statistical Significance Testing in Biomedical Research. *European Journal of Epidemology* 25(Mar.): 225-230.

Sterling, T. D., W. L. Rosenbaum, and J. J. Weinkam. 1995. Publication Decisions Revisited: The Effect of the Outcome of Statistical Tests on the Decision to Publish and Vice Versa. *American Statistician* 49(Feb.): 108-112.

Stekler, H. O. 2007. Significance Tests Harm Progress in Forecasting: Comment. *International Journal of Forecasting* 23: 329-330.

Trejo, Stephen. 1991. The Effects of Overtime Pay Regulation on Worker Compensation. *American Economic Review* 81(Sept.): 719-740.

Wainer, Howard. 1999. One Cheer for Null Hypothesis Significance Testing. *Psychological Methods* 4(2): 212-213.

White, William. 1967. The Trustworthiness of "Reliable" Econometric Evidence. *Zeitschrift für Nationalökonomie* 27(Apr.): 19-38.

Wolff, Edward. 1991. Capital Formation and Productivity Convergence Over the Long Term. *American Economic Review* 81(June): 565-579.

Woodbury, Stephen, and Robert Spiegelman. 1987. Bonuses to Workers and Employers to Reduce Unemployment: A Randomized Trial in Illinois. *American Economic Review* 77(Sept.): 513-530.

Wooldridge, Jeffrey. 2004. Statistical Significance is Okay, Too: Comments on "Size Matters." *Journal of Socio-Economics* 33(Nov.): 577-579.

Zellner, Arnold. 2004. To Test or Not to Test, and If So, How? Comments on "Size Matters." *Journal of Socio-Economics* 33(Nov.): 581-586.

Ziliak, Stephen T., and D. N. McCloskey. 2004. Size Matters: The Standard Error of Regressions in the *American Economic Review*. *Journal of Socio-Economics* 33(Nov.): 527-546. (This article was also published, with permission of the journal just cited, in *Econ Journal Watch* 1(2): 331-358 (**link**).)

Ziliak, Stephen T., and D. N. McCloskey. 2008a. Science Is Judgment, Not Only Calculation: A Reply to Aris Spanos's Review of *The Cult of Statistical Significance*. *Erasmus Journal of Philosophy and Economics* 1(Autumn): 165-170. **Link**

Ziliak, Stephen T., and D. N. McCloskey. 2008b. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice and Lives*. Ann Arbor: University of Michigan Press.

About the Author

Thomas Mayer is emeritus professor of economics, University of California, Davis. His main fields are the methodology of economics and monetary policy. His latest book is *Invitation to Economics*. His e-mail address is tommayer@lmi.net.

Deirdre McCloskey and Stephen Ziliak's reply to this article
Go to Archive of Economics in Practice section
Go to September 2012 issue

EJW
ECON JOURNAL WATCH
Scholarly Comments on Academic Economics

ECON JOURNAL WATCH 9(3)
September 2012: 298-308

Statistical Significance in the New Tom and the Old Tom: A Reply to Thomas Mayer

Deirdre N. McCloskey[1] and Stephen T. Ziliak[2]

LINK TO ABSTRACT

Tom Mayer, an old friend, wrote in 1980 a pioneering paper making the point we make in *The Cult of Statistical Significance*, namely, that testing for fit is not the same thing as testing for oomph. Tom seems now to have reverted to an erroneous, pre-1980 understanding. The direction of movement from Old Tom to New Tom is unusual. Normally speaking once a man grasps the point that Type I error in the absence of a loss function calculated at R. A. Fisher's conventional level of 5% is neither necessary nor sufficient for scientific and commercial discoveries, he does not lose his grip on it.

The economist Lawrence Officer for instance declares that grasping the insignificance of "Fisher significance" has changed his life (Officer 2010). Now instead of meaningless tests of "significance" he can test the actual hypothesis by doing what most scientists do—by looking at magnitude, by looking at the divergence of evidence from the going rhetoric of the field, that is, by doing what scientists call the test of "interocular trauma." Chemists and geologists, most biologists and historians, and almost all physicists test a hypothesis, whether major or subsidiary, by asking if the difference in magnitude between what they see and what they expect hits them between the eyes. Honest, that's what they do. Most sciences rarely use tests of statistical significance. Funny, then, that we economists are addicted to them.

1. University of Illinois at Chicago, 60607.
2. Roosevelt University, Chicago, IL 60605.

Tom Mayer (2012, abs. and 259) suggests that there has been a "debate" in economics about this old and unoriginal point, that fit is not the same thing as substantive importance. Oh, no, we don't think there has been a debate. The orthodox, who very much want to go on using their statistical training in a mechanical way, will *not* debate. When challenged they revert to a sullen silence.

From the beginning of modern statistics many of its theorists and some of its practitioners have been making 'our' point. Yet repeatedly, despite the devastating implications for the orthodox procedures (such as David Hendry's of "test, test, and test" (1980, 403)), the point has been walked past. No: strode past, with a sneer or a shudder or silence.

In economics the sole exception before New Tom to the defensive silence was a paper by Kevin Hoover and Mark Siegler in 2008, published just before our book came out (Hoover and Siegler 2008a). Kevin and Mark had, like Tom, the scientific integrity to *try* defending the routine of "testing" how important a variable is by examining the sampling variance of its fitted β. Yet Kevin and Mark, like Tom, early in their paper admit our point. True, they proceed to call it names ("jejune," for example). But they say outright that they agree with our point. Unhappily, as we noted in our reply, they don't actually understand it (McCloskey and Ziliak 2008). Sadly now we have to give the same reply to New Tom.

Tom, like Kevin and Mark, focuses on Ziliak and McCloskey. But we showed in the book that our point has been made for nearly a hundred years repeatedly even in the sciences that have come to depend on Fisherian dogma. In economics our critics radically simplify their rhetorical task by attacking the naïfs McCloskey and Ziliak, and giving the silent treatment to the identical point made by Edgeworth, Gosset (aka "Student" of Student's *t*), Egon Pearson, Neyman, Jeffreys, Borel, Wald, Yule, Deming, Savage, de Finetti, Good, Lindley, Feynman, Lehmann, DeGroot, Raiffa, Arrow, Milton Friedman, Mosteller, Tukey, Kruskal, Mandelbrot, Wallis, Roberts, Freedman, Meehl, Cohen, Rothman, Leamer, and Zellner, to name but a few.

New Tom's rhetoric throughout is of "a more balanced reading" (2012, abs.). He has come to believe that all sides must have *some* merit (though he gives almost all of the weight to the orthodox side). He is employing, quite in the character of this judicious and fair-minded scholar (true of both Old and New Tom), a Lemma of Compromise. His repeated talk of 'on the one hand, on the other' (as for example in his treatment of Aris Spanos' 2008 review of our work) suggests that he is committed to balance in scholarship. Good. But his version of balance reminds us of a bill before the Indiana Legislature long ago proposing to square the circle and set the value of pi to, for example, precisely 3.00000, *in order to make calculation easier* (and some commentators believe) to honor the rationality of the Holy Trinity (Singmaster 1985). If pi is actually an irrational number estimated

roughly at 3.14159, then for most practical purposes (such as wheel turning) it is *not* a good idea to compromise on, say, 3.07080, halfway between 3.00000 and the correct approximation, because that is the social convention and is easier. Tom remarks mildly that "it is unlikely that either side is totally wrong" (2012, 259). We honor the tone. But in science we try to arrive at mutually exclusive positions and then decide which position is better. Social grace is not the same thing as good science, and square wheels can't carry us to the market. Effect size matters all the way down.

Tom's devotion to balance leads him to quote charges against us without troubling to read the historical evidence, or to think through the charges, or to make his own judgment on the evidence and logic involved. It's like the journalist who is told by his editor to *always* get both sides of the story. There are *always* two sides (or more). Fine. It is one reason Tom edges closer to 3.00000 than to 3.14159. An example among many scattered through the paper is giving credence to "a referee's charge that when Z-M complain about the confusion of statistical with substantive significance they merely reiterate what many psychologists and statisticians have said for several decades" (259). The referee had it seems not read our book. No worries: we do not insist that the whole world read it. But Tom *is* supposed to have read it. Roughly a hundred and fifty of the book's pages, after all, cite, quote, praise, admire, amend, extend, and celebrate "what many psychologists and statisticians have said for several decades" (closer, by the way, to ten decades).[3] Tom cites Hoover and Siegler (2008a, 2008b) with approval on some fifteen occasions. He cites our crushing reply to Hoover and Siegler twice only, once to quarrel with it. Of the thirty or so positive reviews of *The Cult*, ranging from approving notices to raves, in journals from *Nature* to the *Notices of the American Mathematical Society*, Tom cites not one—not even the long and erudite retrospective of our work by Saul Hymans, Director Emeritus of the National Bureau of Economic Research, published in the *Journal of Economic Literature* (Hymans 2009). New Tom does not assemble much of the evidence for a balanced perspective.

Tom repeatedly praises Spanos (2008) for his complaint that we do not complain enough about the *other* things wrong with significance tests, such as specification error. When one of us presented a talk to the Harvard graduate students trying to get across our much more elementary point (Thomas Schelling described it in a blurb for our book as "this very elementary, very correct, very important argument"), an influential economist replied, "Yeah, sure. But *another* serious error is such-and-such," mentioning some fashionable point of textbook econometrics. We don't think such an attitude is a good one. If an engineer uses 3.00000 rather than 3.14159 in practical applications of pi, she is going to get wrong

3. Ziliak and McCloskey (2008b, 1-22, 123-237, 265-287).

results—for some purposes not *too* bad, for others, disastrous. One should get the elementary concepts of our science right before rushing off to apply still another regression technique to confounded data.

Here's "elementary." Tom (2012, 264) speaks of Milton Friedman's great book on the consumption function as finding that "at any given income level farm families have a lower marginal propensity to consume than urban families." But "lower" is not a feature of the numbers themselves. It is always and everywhere an economic and human matter of how low is low. Similarly, Tom imagines that one might hypothesize that as "the expected future price of drugs rises, current drug consumption falls. And you find that it does" (264). But "it does" is always and everywhere a question of how big is big. You have to say *how much* rise of price causes how much fall in consumption, along a scale of big and small meaningful for some human question at issue. The question is not to be answered *inside the numbers themselves*, without a scale. If you ask whether the outside temperature is hot today you are supplying implicitly some standard, some scale meaningful for a human question. Hot for golfing, hot for September, hot for interstellar gas.

Like everyone who makes the "significance" mistake, Tom shifts immediately to the *other* question—to the amount of sampling error on the estimate of a coefficient. Statistical economists after Lawrence Klein first started the routine in economics are always shifting immediately to the question of the sampling error on the estimate of a coefficient, because it gives them a mechanical test, a yes/no or on/off switch, though a test not answering the scientific or policy or other human question at issue: the question of how much. Tom says, "If someone, by adding an additional variable develops a variant that does predict well, this will be of interest to many economists, both to those who want to predict future rates, and those who wonder why the standard theory predicts badly, regardless of the oomph of that variable" (Mayer 2012, 264). But predicting "well" and predicting "badly" are themselves matters of oomph, not fit. If the additional variable said to be relevant to explaining the term structure of interest rates gave us a better explanation in the magnitude of one hundredth of a basis point, the operator of a super millisecond arbitrage engine might care, but the Federal Reserve Board would not. If someone came up with a very large sample that showed the additional variable to be nonetheless "significant" at the 5% level (setting aside power, sample biases, misspecification, and so forth), the other economists would judge her to be naïve—*if* it had occurred to them already that good fit is, after all, not the same thing as quantitative relevance to how big is big. Our point is that it has *not* on the whole occurred to most economists, who are over-trained in Fisherian econometrics and under-trained in the economic approach to uncertainty originated by William Sealy Gosset at the Guinness Brewery in 1904.

The point is made vividly in a recent interocular experiment by Emre Soyer and Robin Hogarth (2012) who tested the forecasting abilities of 257 well-published econometricians (as discussed by Ziliak 2012). When the econometricians were given conventional output such as *t*-tests and R^2s, over 70% of the predictions were wrong (again, by some human scale of mattering). When the econometricians were shown only a scatterplot of data relative to a theoretical regression line, 3% of the predictions were wrong. New Tom needs to hear about the Soyer-Hogarth experiment.

Tom attacks our thought experiment (Mayer 2012, 261, concerning Ziliak and McCloskey 2008b, 23-25, 43-44) about a choice of pills for mother's weight loss, pill Precision versus pill Oomph, writing that "it [the experiment] assumes that we already know the means for the two pills, and it thereby bypasses the need to ask the very question that significance tests address." Wait a minute. Mom and we and New Tom do know for each variety of pill the sample mean, which is after all an unbiased estimate of the population mean. We stipulated the fact, and stipulation or not a sample mean stands for what Mom needs to know. Our thought experiment presupposes large enough samples and good enough experiments to be confident in taking the sample mean to be the best available evidence on the population mean. So knowing the means is not the issue. Yet Tom, driven by his admirable desire to compromise, wants the level of significance ($p < .05$) to decide the choice of weight loss pill. He very much wants to find a middle ground between us and, say, Hoover and Siegler, and more generally the conventional practices of economists. So as usual in the literature he slides over into the *different* question—sometimes a relevant question, usually in science not the main one—of how *certain* we are of the estimated means (modulo sampling theory alone, as though sampling variation were the only source of what Gosset called "real error"; see Ziliak 2011). "Given the existence of sampling error," Tom writes, "how confident can we be about these point estimates?" (2012, 261). He writes again, "Statistical significance is needed to justify treating the coefficient … as a sufficiently reliable stand-in for the true coefficient" (261). But in our example of pill Precision versus pill Oomph, your mother—whose goal is to lose weight, not to publish in economics journals dominated by a mistaken routine for science—is given the best available knowledge about variances, too. She knows the means *and* the variances. She has to choose and, if she is wise, she will use oomph to guide her to the best choice, not the probability of a Type I error in the absence of a loss function. The loss function is weight lost, not satisfaction of a referee with a feeble grasp of statistical theory. To say it yet again: *there are two separate questions*—one question is about oomph, the magnitudes of which Mom and most scientists seek. The other question is about the error terms around such magnitudes, which Mom

has already assessed in the two pills, and is anyway irrelevant to her goal of losing weight.

And, more deeply, how would you know how "true" something is without some standard of truth *within the realm of human importance, beyond the numbers*? That is, how would you know what advice to give as to degrees of belief *without a loss function*? As the Old Tom writes (he occasionally pops up here), "What we need, but do not have, is an explicit loss function" (Mayer 2012, 273). Precisely. That is what we note in numerous examples throughout *The Cult of Statistical Significance*.

Even in gloriously second or third moments, that is, numbers do not contain their own interpretation. Our handful of critics in economics, joined now by New Tom, want technical tricks with the numbers to substitute for scientific considerations of how big is big. Tom thinks that "there is nothing wrong with using asterisks to draw attention to those hypotheses that have passed a severe test" (262), as though sampling error is the chief test that a hypothesis must pass. It's not. Consider. Your doctor has in her possession a conventional regression of health status on two variables, height and weight. Though the standard error on height is delightfully low, that on weight is embarrassingly high, not passing what Mayer (following Deborah Mayo) calls a "severe test" (Mayer 2012, 262). So when you go for a checkup your doctor says, "Your problem is not that you're too heavy, it's that you're too short."

On the basis of a fleeting reference to a paper by our former colleague Joel Horowitz (who well understands the problem, and teaches 'our' point to his students), Tom asserts that tests of statistical significance "are at home in the physical sciences" (Mayer 2012, 256). That is quite wrong. Tom must not have looked into the matter empirically. We realize that economists will be surprised to hear the news, but we say again: physicists and chemists and geologists almost never use such tests. A very minor use is restricted, as Tom (262) concedes, to informing readers about the sampling variation in a measurement of, say, the speed of light, but never as a test of the scientific hypothesis in question, such as that nothing can move faster. If you do not believe us, wander over some afternoon to, say, the physics department and spend an hour or two paging through any of the four versions of *The Physical Review*. You will not have to understand the physics (thank heaven) or the mind-boggling mathematics (double thanks) to notice that the *t*-tests that proliferate in economics journals and the *p* values in psychological journals are not in evidence. The lack of asterisks and R^2s is not because physical scientists do not face samples similar to those economists face (though the physical scientists do exhibit a clearer understanding that a sample must be a sample, not a universe; economists are frequently found insisting on the mathematics of sampling theory when they have the urn of nature spilled out before them). Economists in their modest way usually assume that people who don't use our

wonderfully sophisticated econometric methods of testing (no one, by the way, including us, is complaining about *estimation*, which is very useful compression of data) must be sadly lacking in the math to do so. But the hypothesis won't survive an hour or two with *The Physical Review*.

Tom misunderstands our distinction between philosophical existence questions and scientific magnitude questions. He cites (262) for example the mistaken assertion by Hoover and Siegler that the 1919 debate in physics about the bending of light around the sun was a matter of existence, not magnitude. On the contrary, within the limits of exactitude in the instrumentation it was a question of magnitude, as anyone knows who has actually read the literature on the matter (see, for example, Jeffreys 1919). Without an *important* bending, Einstein fails (and his exact prediction did fail for some years until a bias in the instruments was cleared up). Triviality is a matter of oomphility. It is not merely some property of the numbers themselves, independent of a scale of judgment along which to measure it.

Philosophy, theology, and mathematics care only about existence, not magnitude. But none of those admirable departments of the intellect is an empirical science. In the Department of Mathematics if *just one* even number is found that is not the sum of two primes, Goldbach's Conjecture will be discarded forever. The Conjecture remains unproven, in the mathematician's sense of the term. Yet it has been shown to be true by calculation up to very, very large numbers. For engineering or physics the numbers are large enough to treat the Conjecture as true—say, for purposes of making a computer lock. The trouble is that most economists have learned their math in the Department of Mathematics, not the Department of Engineering. So they think that in an empirical science you can test for existence separately from magnitude. They give the usual, philosophically naïve justification that "it makes little sense for scientists to try to measure the size of something that does not exist" (Mayer 2012, 262). That's not how actual science works. In actual science one wants to know how big is big, every time. There's no use for an existence of infinitesimal size ε. Not in the vast bulk of science at any rate (that is, in $[1 - \varepsilon][100]$ percent of it).

We say reluctantly that we find New Tom's small sample finding that economists are in fact *not* misled by mistaking statistical for substantive significance, … well, incredible. Surveys of the sort he and we did of *AER* papers involve judgment calls, and we suppose he reckons that if half of the profession gets it (which he claims based on his little sample claiming to re-test our AER studies), that's good enough. Our own experience—and the experience of by now hundreds of critics of null hypothesis significance testing in dozens of fields of science—is that we find the mistake in eight or nine out of every ten empirical papers in economics and other life and human sciences, from medicine to psychology.

We're glad that Tom brings up court cases. He argues that:

> As long as the racial variable has the expected sign and is significant, you have bolstered a claim of racial discrimination. By contrast, suppose you find a substantial oomph for the racial variable, but its *t* is only 1.0. Then you do not have as strong a case to take to court. Z-M might object that this merely shows that courts allow themselves to be tricked by significance tests, but don't courts have to consider some probability of error in rendering a verdict? (Mayer 2012, 263)

Well, it turns out that a court, the Supreme Court of the United States, has weighed in on the question. In "Brief of *Amici Curiae* Statistics Experts Professors Deirdre N. McCloskey and Stephen T. Ziliak in Support of Respondents" (2010) before the U.S. Supreme Court, an eminent New York law firm drew upon our book and our articles to argue an important case of biomedical and securities law. In March 2011 the U.S. Supreme Court decided in 'our' favor—nine to zero (*Matrixx Initiatives, Inc. v. Siracusano* 2011). Statistical significance, the Court decided, is *not* necessary to prove biomedical, investor, and economic importance. It is the law of the land, and should be of econometrics.

Tom is right that the "debate" is unlikely to end unless people stop the silent treatment or the angry shouting, and start actually listening to each other. As the novelist and philosopher Iris Murdoch wisely put it, "Humility is not a peculiar habit of self-effacement, rather like having an inaudible voice, [but] it is selfless respect for reality and one of the most difficult and central of all virtues" (Murdoch 1967, 95). A question, then, in all humility has to be answered:

Is statistical significance a wise and wonderful tool with which economic science has made great strides? Your local econometrician will affirm that it is, and will press the graduate committee to insist on still more *t*-testing econometrics, instead of educating the students in quantitative methods used in progressive sciences such as engineering, physics, scientific brewing, agriculture, cosmology, geology, and history—the methods of experimentation, simulation, surveys, narrative, accounting, canny observation, interocular trauma, Bayesian decision analysis, the comparative method, natural experiments, primary documents, and the economic approach to the logic of uncertainty, lacking in contemporary econometrics.

Econometricians have been claiming proudly since World War II that significance testing is the empirical side of economics (most youngsters in economics think that "empirical"—from the Greek word for "experience"—simply *means* "collect enough data to do a significance test"). Tjalling Koopmans's influential book of 1957, *Three Essays on the State of Economic Science*, solidified the claim. But if you take the con out of econometrics (and the "trics," and the "me" too) you are not left with much (actually, just the *cri de cœur* "eo!"). What major scientific

issue since the War has been decided by tests of statistical significance? Yes, we understand: your *own* view by your *own* tests of monetarism or the minimum wage or whatever. Then why don't those misled other economists agree with you? We said "decided."

Geologists decided in the 1960s that plate tectonics was correct, after resisting it for fifty years. Historians decided in the 1970s that American slavery was profitable in a pecuniary sense, after resisting that claim for a hundred years. Mayan archaeologists decided in the 1980s that Mayan script was not mainly ideographic, after resisting for forty years. Physicists decided in the 1990s that most of matter and energy in the universe was "dark," after declaring for decades that they were close, so, so close, to a Theory of Everything, and needed only some hundreds of billions of dollars to get to it. These are examples of actual progress in science. If the rhetoric of significance made any sense, all these advances could have been based on the level of statistical significance. None actually were. And so too in economics.

The loss-functionless test of statistical significance is a tool of stagnant anti-progress in economics and a few other sciences. Let's bury it, and get on to empirical work that actually changes minds.

References

Hendry, David F. 1980. Econometrics—Alchemy or Science? *Economica* 47: 387-406.

Hoover, Kevin, and Mark Siegler. 2008a. Sound and Fury: McCloskey and Significance Testing in Economics. *Journal of Economic Methodology* 15(Mar.): 1-38.

Hoover, Kevin, and Mark Siegler. 2008b. The Rhetoric of "Signifying Nothing": A Rejoinder to Ziliak and McCloskey. *Journal of Economic Methodology* 15(Mar.): 57-68.

Hymans, Saul. 2009. Review of *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice and Lives*, by Stephen T. Ziliak and Deirdre N. McCloskey. *Journal of Economic Literature* 47(2): 499-503.

Jeffreys, Harold. 1919. On the Crucial Test of Einstein's Theory of Gravitation. *Monthly Notices of the Royal Astronomical Society* 80(Dec.): 138-154.

Koopmans, Tjalling. 1957. *Three Essays on the State of Economic Science*. New York: McGraw Hill.

Mayer, Thomas. 1980. Economics as an Exact Science: Realistic Goal or Wishful Thinking? *Economic Inquiry* 18(Apr.): 165-178.

Mayer, Thomas. 2012. Ziliak and McCloskey's Criticisms of Significance Tests: An Assessment. *Econ Journal Watch* 9(3): 256-297. **Link**

McCloskey, Deirdre N., and Stephen T. Ziliak. 2008. Signifying Nothing: Reply to Hoover and Siegler. *Journal of Economic Methodology* 15(Mar.): 39-56.

McCloskey, Deirdre N., and Stephen T. Ziliak. 2010. *Brief of* Amici Curiae *Statistics Experts Professors Deirdre N. McCloskey and Stephen T. Ziliak in Support of Respondents*, ed. Edward Labaton, Ira A. Schochet, and Christopher J. McDonald. No. 09-1156, Matrixx Initiatives, Inc., et al. v. James Siracusano and NECA-IBEW Pension Fund. November 12. Washington, D.C.: Supreme Court of the United States.

Murdoch, Iris. 2001 [1967]. *The Sovereignty of Good*. London: Routledge.

Officer, Lawrence. 2010. Personal communication, University of Illinois-Chicago.

Singmaster, David. 1985. The Legal Values of Pi. *Mathematical Intelligencer* 7: 69-72.

Soyer, Emre, and Robin Hogarth. 2012. The Illusion of Predictability: How Regression Statistics Mislead Experts. *International Journal of Forecasting* 28(3): 695-711.

Spanos, Aris. 2008. Review of Stephen T. Ziliak and Deirdre N. McCloskey's *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice and Lives*. *Erasmus Journal for Philosophy and Economics* 1(Autumn): 154-164. **Link**

Ziliak, Stephen T. 2011. W. S. Gosset and Some Neglected Concepts in Experimental Statistics: Guinnessometrics II. *Journal of Wine Economics* 6(2): 252-277.

Ziliak, Stephen T. 2012. Visualizing Uncertainty: Is a Picture Worth a Thousand Regressions? *Significance* (Royal Statistical Society) 9(5): forthcoming.

Ziliak, Stephen, and D. N. McCloskey. 2008a. Science Is Judgment, Not Only Calculation: A Reply to Aris Spanos's Review of *The Cult of Statistical Significance*. *Erasmus Journal of Philosophy and Economics* 1(Autumn): 165-170.

Ziliak, Stephen, and D. N. McCloskey. 2008b. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice and Lives*. Ann Arbor: University of Michigan Press.

Cases Cited

Matrixx Initiatives, Inc., et al. v. Siracusano, et al., 563 U.S. ___ (2011).

About the Authors

Deirdre N. McCloskey is UIC Distinguished Professor of Economics, History, English, and Communication, University of Illinois at Chicago, and Professor of Economic History, University of Gothenburg, Sweden. For more information about her books, articles, essays, and interviews, visit her website and blog at **http://deirdremccloskey.org**. Her email address is deirdre2@uic.edu.

Stephen T. Ziliak is Trustee and Professor of Economics at Roosevelt University Chicago. For more information about his books, articles, essays, and interviews, visit his websites at **http://sites.roosevelt.edu/sziliak** and **http://stephentziliak.com**. His email address is sziliak@roosevelt.edu.

Go to Archive of Economics in Practice section
Go to September 2012 issue

EJW
ECON JOURNAL WATCH
Scholarly Comments on Academic Economics

ECON JOURNAL WATCH 9(3)
September 2012: 309-349

Mankiw vs. DeLong and Krugman on the CEA's Real GDP Forecasts in Early 2009: What Might a Time Series Econometrician Have Said?

David O. Cushman[1]

LINK TO ABSTRACT

Introduction

At the end of February 2009, the Council of Economic Advisers (CEA) of the new Obama administration forecasted a strong rebound in the U.S. economy from the recession. The CEA predicted that, after a further dip in 2009, real GDP would recover strongly, growing at annual rates of over 4% in 2011 and 2012 and achieving cumulative growth of 15.6% by 2013 compared to 2008. The CEA based its forecast on the newly decided size of the fiscal stimulus and on the "key fact…that recessions are followed by rebounds" and "deeper recessions are typically followed by more rapid growth."[2]

A few days later, Greg Mankiw expressed doubts in a blog entry (Mankiw 2009b). He suggested that the administration was "premising its forecast on the economy being *trend stationary*." If so, shocks have only temporary effects on real GDP. After a large negative shock and its resulting recession, real GDP would

1. Westminster College, New Wilmington, PA 16172.

2. The CEA gave its rebound discussion and forecasts for 2009 and 2010 in "Economic Projections and the Budget Outlook" (CEA 2009). The full set of administration forecasts through 2019 is in Office of Management and Budget (2009), Table S-8, issued at the same time and cited in CEA (2009).

rebound to its unaltered long-run growth path, growing at a higher-than-normal rate to get there. Mankiw quoted the abstract of a paper he coauthored with John Campbell (Campbell and Mankiw 1987a), in which they noted they were "skeptical of this implication" (that shocks do not affect real GDP's long-run growth path) for post-war U.S. real GDP. They argued instead for the unit root hypothesis. "It contrasts starkly with the trend-stationary hypothesis," wrote Mankiw in his blog (2009b). If real GDP contains a unit root, shocks tend to permanently change real GDP's growth path. In fact, Campbell and Mankiw (1987a) had concluded that a 1% negative shock to real GDP would lead to a permanent reduction in the growth path of *even more than* 1%. After a recession, one would therefore expect no rebound in real GDP.

Brad DeLong, in his own blog (DeLong 2009), immediately retorted that one needed to distinguish between permanent and transitory effects.

> A fall in production that does not also change the unemployment rate will in all likelihood be permanent. A fall in production that is accompanied by a big rise in the unemployment rate will in all likelihood be reversed. You have to do a bivariate analysis—to look at two variables, output and unemployment.

The 2008 decline in real GDP was, as all know, accompanied by an increase in the unemployment rate, and so there would be a rebound in real GDP, just as predicted by the CEA. "And that is certainly the way to bet," concluded DeLong.

Paul Krugman quickly took up the argument in his blog (Krugman 2009). In an entry provocatively titled "Roots of evil (wonkish)," Krugman wrote:

> I always thought the unit root thing involved a bit of deliberate obtuseness—it involved pretending that you didn't know the difference between, say, low GDP growth due to a productivity slowdown like the one that happened from 1973 to 1995, on one side, and low GDP growth due to a severe recession. For one thing is very clear: variables that measure the use of resources, like unemployment or capacity utilization, do NOT have unit roots: when unemployment is high, it tends to fall. And together with Okun's law, this says that yes, it is right to expect high growth in future if the economy is depressed now.

Finally, Mankiw (2009c) wrote:

> Paul Krugman suggests that my skepticism about the administration's growth forecast over the next few years is somehow "evil." Well, Paul, if you are so confident in this forecast, would you like to place a wager

> on it and take advantage of my wickedness? Team Obama says that real GDP in 2013 will be 15.6 percent above real GDP in 2008. (That number comes from compounding their predicted growth rates for these five years.) So, Paul, are you willing to wager that the economy will meet or exceed this benchmark? I am not much of a gambler, but that is a bet I would be happy to take the other side of (even as I hope to lose, for the sake of the economy).

Krugman made no response to this that I know of, but given DeLong's certainly-the-way-to-bet statement above, Mankiw's bet may have been directed to the wrong person.[3]

Now, this blog exchange was more interesting than most because it involved three of what may be the four most popular economics blogs (Davis et al. 2011). Moreover, at the time I was immediately sympathetic to Mankiw. I liked Mankiw based on his style of presenting ideas, on having met him, and on having had a few, brief (and pleasant) email exchanges (there have been more since then). And I am also basically a conservative, market-favoring economist, just as Mankiw seems to be. In contrast, I am not a fan of shoot-from-the-hip analysis accompanied by incomplete evidence and, yet, extreme self-confidence; in my opinion, such often appears in DeLong's and Krugman's blogs. Nor do I like their frequent name calling, condescension, and snark. Thus, I was interested in trying to prove them wrong.[4] Yet, their point about accounting for permanent versus transitory shocks in forecasting was a good one.[5]

Therefore, I wondered what a careful, non-DeLong/Krugman forecasting analysis would indicate. Would it confirm that the CEA forecast was an obvious one, as DeLong and Krugman expressed in their quick retorts? Or would Mankiw's bet look pretty good?

Other forecasts were then available. An Office of Management and Budget (OMB) report (2009) referenced by the CEA, which Mankiw likely looked at to

3. Although Mankiw's central source of rebound skepticism certainly seemed to be the unit-root idea, his initial blog on this (Mankiw 2009b) presented two additional reasons for doubt. For the interested reader I elaborate and clarify them in the Appendix.

4. In the interest of fuller disclosure, I'll reveal that more recently, in September 2011, I received (in contrast to pleasant Mankiw emails) an unpleasant email from DeLong. It was in response to a comment I had attempted to post on his blog. He wrote me, "Shame on you for trying to confuse the issue." He also did not allow the comment to be posted. As my project was already well underway, the event was clearly not an initial motivator, but may have served to spur me on. In case the reader wants to assess whether I was "confusing the issue," I provide details in the Appendix. The episode also provides a case study of the sorts of things I don't like in DeLong's and Krugman's blogs.

5. In his "Wanna Bet" blog, Mankiw (2009c) pointed out that Campbell and he (1987b) had, in fact, investigated transitory versus permanent shocks in GDP. The paper's abstract concludes, "We find no evidence for the view that business cycle fluctuations are more quickly trend-reverting."

get the CEA growth forecast through 2013, presented not only the CEA forecasts but also the forecast of the Congressional Budget Office (CBO) from January and the Blue Chip consensus forecast from February.[6] In contrast to the CEA forecast of 15.6% growth from 2008 to 2013, the CBO forecast was for 12.4% and the Blue Chip consensus for just 9.1%. A February revision by the CBO (Elmendorf 2009), which took account of the stimulus bills then being considered by Congress, implied slightly higher growth of 12.7% by 2013.[7] Since Mankiw's bet was that he would lose only if growth met or exceeded 15.6%, it seems he was not being all that bold (in line with being "not much of a gambler"!), independent of his unit root story.

In any event, I had no idea (nor would any outsider, I presume) how the CBO and Blue Chip forecasts were constructed. Therefore, it was unclear to me to what extent the CBO and Blue Chip forecasts reflected the issues that Mankiw, DeLong, and Krugman deemed important.

I decided to perform my own analysis. It would be based on forecasting techniques that seemed to me to be standard in the sense of being found in many standard econometric and time series textbooks and used often in the journal literature. I would also add a few refinements from the recent journal literature. I wanted the analysis to be reasonably thorough and as unbiased as I could manage.

I had other projects to complete, however, and two years passed before I returned to the idea. By then it was beginning to appear that Mankiw would have been on the way toward winning his bet. Real GDP for 2010 (as of the February 25, 2011, Bureau of Economic Analysis (BEA) release) was a mere 0.1% higher than in 2008, not 2.0% higher as forecasted by the CEA. And at the time of final revisions to this paper (August 2012), a Mankiw win seems even more likely, with real GDP for 2011 (as of the July 27, 2012, BEA release) only 1.0% higher than in 2008 instead of the CEA's forecasted 6.0%. But such developments shed no light on the outcome of the project I imagined in 2009.

---

6. The Blue Chip forecasts for 2009 and 2010 in OMB (2009) are from February 2009, and the subsequent years from October 2008. The Blue Chip consensus is the average of forecasts by approximately 50 private forecasters of a number of macroeconomic variables. The monthly issues of Blue Chip Economic Indicators present annual forecasts for the next two years and the March and October issues additionally present longer forecasts. See Aspen Publishers' website (**link**).

7. I doubt that Mankiw, DeLong, or Krugman would have taken into account the revised CBO figures as far as growth through 2013 is concerned, because the figures were not presented in a way easy to compare with 2008's calendar-year value. The revised CBO figures were presented as revisions to fourth-quarter levels relative to the CBO's January "baseline" values. Thus, computing the calendar-year values requires some interpolations and other assumptions. Details are in the Appendix. As noted in the previous footnote, the February 2009 Blue Chip forecast for 2013 was actually from October 2008, and so it's not clear whether it accounted for the possibility of stimulus.

In spring 2011, I decided to go ahead with the project. But I realized that the forecasting needed to proceed as if it were still March 2009, because it would not be fair to utilize data, information, or techniques only more recently available. I must also admit that developments since March 2009 have surely affected my motivation to go through with the investigation, but I have striven to maintain the same goal of an unbiased econometric analysis that I had then.

I began by imagining a hypothetical time series econometrician who would apply ARIMA (autoregressive integrated moving-average) models (Box and Jenkins 1970) to first differences of real GDP, as done by Campbell and Mankiw (1987a). ARIMA models are sometimes referred to as ARIMA(p,d,q) models, where the autoregressive lag order is p, the order of differencing is d, and the moving average lag order is q. ARIMA(p,d,q) models have been very popular for forecasting and so I decided to refer to my hypothetical econometrician as PDQ (not to be confused with the infamous classical composer P. D. Q. Bach). PDQ is envisioned as male and I will sometimes refer to PDQ with masculine pronouns (he, him).[8] The popularity of ARIMAs for forecasting is attested to by their appearance in this capacity in many well-known textbooks that PDQ would know of (such as Pindyck and Rubenfeld 1997, Enders 2004, Tsay 2005, and Diebold 2008). Graham Elliott and Allan Timmerman (2008, 23) discuss the ARIMA model's "historical domination" in forecasting. Moreover, PDQ has seen ARIMA models applied in the literature to analyze the time series properties of real GDP. Prominent examples are Campbell and Mankiw (1987a), alluded to in Mankiw's blog post, and James Morley, Charles Nelson, and Eric Zivit (2003).

But as I proceeded, I realized that ARIMA modeling did not obviously address DeLong and Krugman's point about using the unemployment rate to help make the forecast. To do so in an obvious way, I felt PDQ should do what DeLong (2009) had suggested: employ a bivariate approach. DeLong's blog post (2009) included a graph of historical real GDP growth rates plotted against earlier unemployment rates. A regression line was included. It showed high unemployment being followed by higher than normal economic growth. But DeLong gave no other statistical results for the regression. Mankiw (2009c) questioned the statistical significance of the line in what he called a "cloud of points." Thus, I wanted PDQ to follow the bivariate suggestion, but with something more credible and *a lot* more thorough.

The obvious answer to me was to add a bivariate VAR (vector autoregression) approach to the project. Like ARIMAs, VARs are popular for forecasting. For example, in his well-known textbook, William Greene (2003, 587) writes that for "forecasting macroeconomic activity…researchers have found that

8. PDQ can also devote full time to the project, and can thus finish it much faster than I have been able to.

simple, small-scale VARs without a possibly flawed theoretical foundation have proved as good or better than large-scale structural equation systems." Similarly, in their abstract Todd Clark and Michael McCracken (2006) write, "Small-scale VARs are widely used in macroeconomics for forecasting U.S. output, prices, and interest rates." The paper goes on to reference empirical papers that have done so. Forecasting with VARs is discussed in a number of widely used texts (e.g., Enders 2004, Lütkepohl 2005, Stock and Watson 2007, and Diebold 2008).

PDQ's main points and findings

I am now going to summarize PDQ's key points, procedures, and forecasts. To document PDQ's attempt to be thorough, I give a more detailed, blow-by-blow account, with graphs of the forecasts, in the subsequent sections of the paper. Still more details are found in the occasionally referenced Appendix. Here's the summary:

1. Real GDP can have a unit root and there can nevertheless be rebounds without necessarily implying a trend-stationary process, contrary to what some might infer from Mankiw's blog entries.
2. Though univariate, ARIMAs allow for both permanent and transitory effects. In contrast to DeLong's (2009) assertion, you do not necessarily need a bivariate approach. However, *if* unemployment rate fluctuations capture most of the transitory effects, a bivariate output-unemployment VAR may be better.
3. Because underlying shocks are usually unobserved and output and unemployment are simultaneously determined, neither an ARIMA nor a VAR model can fully identify transitory and permanent shocks. This counters Krugman's (2009) claim that it is easy to know whether the source of a GDP slowdown is transitory or permanent.
4. Both the ARIMA and the VAR approaches require lag order choices. But rather than pick one set of lag orders for each estimation approach, PDQ combines the forecasts of many lag order models using a recent approach called model averaging.[9] PDQ tries both AIC (Akaike) and BIC (Schwarz-Bayesian) model weights in the averaging.
5. PDQ conducts structural stability tests. They suggest that forecast model estimation should start in 1986:3, rather than earlier, to lessen possible misspecification bias in the forecasts.

9. The Blue Chip consensus is a form of model averaging; it is private forecaster averaging.

6. PDQ computes model-averaged forecasts from the ARIMAs and VARs, reported later in the paper. He goes on to compute overall AIC and BIC average forecasts for his ARIMA and VAR forecasts. The resulting forecasted 2008-2013 growth rates are:
    - ARIMA/VAR (AIC): 13.8%
    - ARIMA/VAR (BIC): 11.5%
7. The VAR model does not dominate the weights, contradicting DeLong's (2009) assertion that "you have to…look at two variables, output and unemployment."
8. PDQ's overall forecasts for 2008-2013 growth straddle the CBO's 12.4% and revised 12.7% forecasts. Like the January CBO and Blue Chip consensus forecasts, they are well under 15.6%. Therefore, PDQ thinks *Mankiw would probably win his bet*.
9. PDQ also calculates confidence bands for his forecasts, unlike the CEA, CBO, or Blue Chip consensus.[10] Plus-minus one standard deviation bands (containing roughly 68% of the probability) for PDQ's overall forecasts for 2013 are:
    - ARIMA/VAR (AIC): 10.6% to 17.0%
    - ARIMA/VAR (BIC): 7.8% to 15.3%
10. Using his overall, ARIMA/VAR model-averaged forecast standard errors, PDQ computes Mankiw's probability of losing. It is only 14% (BIC weights) to 28% (AIC weights). Accordingly, the extreme confidence exuded by DeLong and Krugman in the CEA forecast is not warranted.
11. The CEA forecast is very similar to several variations of a trend-stationary forecast, as Mankiw speculated.
12. Does PDQ forecast any rebound at all? With respect to the eventual long-run equilibrium net of trend, none of PDQ's model-averaged forecasts indicate a rebound from 2008's annual figure, and they indicate at best a trivial one from 2008:4.
13. Things that PDQ could not know: Through 2011, the Blue Chip forecast of real GDP is most accurate, followed closely by the January 2009 CBO forecast. But given the most recent value of real GDP, for 2012:2, the best forecasts for the next few years will likely turn out to be the Blue Chip consensus and PDQ's ARIMA forecasts. However,

10. In the case of the Blue Chip consensus, one could examine the range of individual forecasts, but, while useful, this would not have any known relation to probabilities as does a confidence interval. The revised, February CBO forecasts consist of "high" and "low" values, but there is no explanation except that the range "encompasses a majority of economists' views." The 12.4% and 12.7% values are midpoints.

the CBO's forecasted long-run growth rate of only 2.2% by 2018 is already starting to look plausible.

Foundations for two standard forecasting models

The basic idea of the forecasts to be computed by PDQ is to extrapolate from past movements of key variables to predict future movements of real GDP. The details of government and private response to downturns are not modeled. Instead, for example, if governments have successfully responded to recessions in the past, then low GDP generally leads to stronger recoveries than otherwise, and the model will forecast this when GDP is low.

The ARIMA and VAR that PDQ will use can be derived from a basic state-space model:

$$y_t = \mu_{t-1} + \varepsilon_t \;; \quad \varepsilon_t \sim NID(0, \sigma_\varepsilon^2) \tag{1},$$

$$\mu_t = \mu_{t-1} + a + \eta_t \;; \quad \eta_t \sim NID(0, \sigma_\eta^2) \tag{2}.$$

The log of real GDP, y, is determined by a permanent component (μ) and a transitory component (ε). The permanent component is determined by a constant trend (a) and permanent shocks (η). Permanent shocks to capacity or labor force participation (shocks to the long-run growth path) are given by η, and temporary shocks to capacity utilization or employment by ε. The shocks probably cannot be directly observed. But note that the overall shock to y_t consists of a combination of permanent and temporary shocks, and the temporary ones are by definition reversed. Thus, observed y_{t+1} will sometimes show "rebounds," the frequency and size of which will depend on the relative sizes of the temporary and permanent shocks and their correlation. Regardless, y_t has a unit root. Permanent shocks occur every time period.

Ruey Tsay (2005) and Rob Hyndman et al. (2008) show how the equations (1) and (2) have an ARIMA(0,1,1) as a reduced form:

$$\Delta y_t = a + \upsilon_t - \theta_1 \upsilon_{t-1} \tag{3}$$

The value of θ_1 has a relationship to the unobserved shocks in (1) and (2) that depends on certain assumptions. One common assumption (e.g., found in Tsay 2005) is that the permanent and transitory shocks are uncorrelated. In this case, θ_1 is related (nonlinearly) to the signal-to-noise ratio, the relative variances of the

permanent and transitory shocks. The lower is the signal-to-noise ratio, the higher is θ_1. Another common assumption (promoted by Hyndman et al. 2008) is that the shocks are perfectly correlated: $\eta_t = d\varepsilon_t$. In this case, $\theta_1 = 1 - d$ and $\upsilon_t = \varepsilon_t$. In both models, the higher is θ_1, the more important are transitory shocks relative to permanent ones. Various serial correlation processes for the shocks, and other extensions, lead to more general ARIMA(p,d,q) models:

$$\Delta^d y_t = \sum_{k=1}^{p} \varphi_k \Delta^d y_{t-k} + a + \upsilon_t - \sum_{k=1}^{q} \theta_k \upsilon_{t-k} \tag{4}$$

where p and q are the autoregressive and moving average lag orders, and d is the order of differencing to achieve stationarity.

The discussion above indicates that, despite utilizing a univariate model, ARIMA forecasting takes into account both permanent and transitory effects and thus addresses, to a certain extent, DeLong and Krugman's observations that these effects need to be distinguished in forecasting. The distinction (or identification) in the ARIMA is, however, not perfect. The ARIMA forecasting approach cannot identify what portion of any observed shock is permanent or temporary (unless the Hyndman et al. (2008) assumption of perfect correlation is correct). Instead, ARIMA estimation supposes that a constant fraction of a recently observed shock will be reversed. The fraction is based on the estimated average tendency of observed shocks to be transitory. In the basic ARIMA(0,1,1) case, this tendency is related to the θ_1 estimate, and in more complicated ARIMAs to all the φ's and θ's.

PDQ thus pursues the VAR approach to more directly address DeLong and Krugman's criticism. If one is willing to associate unemployment with one or both shocks in the state-space model, then equations (1) and (2) can be transformed into a VAR that can be used for forecasting.

PDQ supposes that, as suggested by DeLong (2009), the transitory shock is related to unemployment: $\varepsilon_t = -b_1(un_t - \overline{un}) + \tau_t$, where un is the unemployment rate, $\overline{un}$ its mean, and τ an independent transitory effect not captured by the unemployment rate. He substitutes this into equation (1). PDQ also allows for the possibility that unemployment fluctuations could be the source of some of the permanent effects: $\eta_{t-1} = -c_1(un_{t-1} - \overline{un}) + \psi_{t-1}$. For example, the skills of the unemployed may deteriorate so that they are less likely to be rehired (Pissarides 1992).

After the substitutions, PDQ transforms equations (1) and (2) into a first difference equation:

$$\Delta y_t = a_1 - b_1 un_t + (b_1 - c_1) un_{t-1} + \tau_t - \tau_{t-1} + \psi_{t-1} \quad (5)$$

where $a_1 = a + c_1 \overline{un}$. Finally, PDQ supposes that fluctuations in *un* are related to Δy_t and an underlying transitory shock ω:

$$un_t = a_2 - b_2 \Delta y_t + \omega_t \quad (6)$$

PDQ notes that Δy and *un* are now simultaneously determined. Equation (5) says that more employment increases the supply of output and equation (6) says that more output increases the demand for labor. Equations (5) and (6) constitute a first-order VAR but cannot be used for forecasting because of the presence of unlagged variables. To get rid of them, PDQ solves the system for the reduced form VAR:

$$\Delta y_t = \left[(a_1 - b_1 a_2) + (b_1 - c_1) un_{t-1} + \tau_t - \tau_{t-1} + \psi_{t-1} - b_1 \omega_t\right] \Big/ (1 - b_1 b_2) \quad (7),$$

$$un_t = \left[(-b_2 a_1 + a_2) - b_2 (b_1 - c_1) un_{t-1} - b_2 (\tau_t - \tau_{t-1} + \psi_{t-1}) + \omega_t\right] \Big/ (1 - b_1 b_2) \quad (8).$$

Neither of the reduced form VAR equations has any explicit lags of Δy. Moreover, the portion of the transitory error ε that is not captured by *un*, which is τ, leads to the composite error terms being equivalent to a moving average process, just as the transitory error in equation (1) leads to a moving average process in (3). In estimation of the VAR, the moving average can be approximated with a sufficient number of lags of Δy and *un*. But if *un* captures most of the transitory effects, the moving average part will be small and additional lag terms trivial in importance. However, serial correlation in ψ and ω will also introduce additional lag terms.[11] The simultaneous determination of Δy and *un* means that it will not be possible to identify or separate out the effects of the underlying transitory and permanent shocks. Contrary to the implications of DeLong and Krugman, observed values of the unemployment rate do not necessarily measure transitory shocks.

PDQ believes that the relative merits of the ARIMA versus the VAR can be summarized as follows. If the unemployment rate in the VAR captures a high

11. The moving average aspect of equations (7) and (8) could be addressed by estimating the system as a VARMA (vector autoregressive moving average) model, just as the first-difference version of equations (1) and (2) can be estimated by an ARMA. However, PDQ has read of the difficulties in the identification and estimation of VARMAs discussed in Lütkepohl's (2005) time series text, where four chapters are devoted to the model. Perhaps because of the difficulties, the JMulti computer program (**link**), which is directly based on Lütkepohl's work, does not include VARMA estimation.

portion of the transitory shocks assumed unobservable in the state-space/ARIMA approach, then the VAR forecasts should be better because the VAR model will have less noise. For example, the transitory shock ω_t will be captured to a significant extent in the un_t value that will be used to forecast y_{t+1} using equation (7). But if *non*-unemployment transitory shocks are important, then the VAR will have a significant moving average error and contain more noise than the ARIMA unless its lag order is rather long (which then reduces statistical efficiency). In contrast, the ARIMA does not need long lag orders to handle an unspecified transitory shock.

Data and preliminary analysis

Data set

PDQ realizes that he should use the same data set as was available to the CEA. The CEA (2009) stated that "[t]he Administration's economic assumptions were largely completed in early January and finalized on February 3rd." Therefore, the CEA presumably had access to the Bureau of Economic Analysis (BEA) release of January 30, 2009, containing data through 2008:4.[12] The bloggers, meanwhile, could have looked at the February 27th release, but they would likely agree with PDQ that assessment of the CEA forecast should be based on the data available to the CEA at the time and not subsequent revisions. For unemployment the CEA would have had access to the January 9th release from the Bureau of Labor Statistics (BLS), and so unemployment rates reported as of this date are what PDQ uses.[13]

PDQ decides he needs to address three issues before actually computing forecasts. (1) What specific ARIMA and VAR lag specifications should he use? (2) His estimation period will end in 2008:4, but when should it begin? The available quarterly data starts soon after World War II, but the literature contains well-known evidence of structural changes since then, which, if included in the forecasting model estimation period, could bias the forecasts. (3) The ARIMA,

12. The archive of releases is available at the BEA's website (link). The specific real GDP series is billions of chained (2000) dollars, seasonally adjusted at annual rates. The values are transformed to logs for estimation and forecasting.

13. The unemployment rate is the seasonally adjusted rate for age 16 and over, series LNS14000000. The current full data set is found at the BLS's website (link), but it reflects revisions unavailable to PDQ. However, each January the BLS generates revised values for the most recent five years. I thank Karen Kosanovich of the BLS for providing me the data released in January 2009. After substituting the January 2009 data for the more recently released data for 2004-2008, we have the complete unemployment series as of January 2009.

equation (4), and the VAR, equations (7) and (8), contain Δy and *un* and thus assume that *y* has a unit root and is first-difference mean stationary, and that *un* is mean stationary. Is this reasonable?

Lag order selection

Many methods exist for choosing a forecasting model when many lag orders and variables are possible. For univariate models, Tsay (2005) and Hyndman et al. (2008) emphasize information criteria such as the AIC and BIC criteria. Hyndman et al. (2008) seem to slightly favor the AIC. Diebold (2008) favors the BIC. For forecasting with VARs, Helmut Lütkepohl (2005) also suggests using information criteria. He seems to have no clear favorite. PDQ decides to use both the AIC and BIC.

PDQ is, however, impressed with the relatively recent approach of using not just the forecasts of the top model from a given criteria, but weighted averages of the forecasts from many models. The weights or probabilities are computed from the models' AIC or BIC values (see Koop and Potter 2003, Hansen 2007, and Wright 2008). Bruce Hansen (2007) favors AIC over BIC based weights.[14] The Appendix gives the formulas.

PDQ uses the same ARIMAs as in Campbell and Mankiw (1987a), 16 models with lag orders *p* and *q* of 0 to 3. In the weighting, PDQ assumes equal priors, a typical approach. For the VAR models, PDQ increases the maximum lag order to 4. Because equations (7) and (8) clearly include the possibility of different lag orders of variables within *and* between equations, PDQ allows Δy and *un* in each equation to each have different lag orders, ranging from 0 to 4 and thus yielding a total of 625 VAR models. Once again, PDQ assumes equal priors.[15]

When to start the estimation period

The full quarterly U.S. real GDP set from the BEA begins in 1947:1, and many papers analyzing the time-series properties of U.S. real GDP have used this starting date. The BLS quarterly unemployment data starts at almost the same point, 1948:1. Therefore, PDQ defines his full data set as starting in 1948:1, with first differences starting in 1948:2. However, PDQ recollects the famous paper by

14. However, Hansen (2007) is specifically promoting the use of the less well-known Mallows criteria over either the AIC or BIC. But this is in a single equation environment. The Mallows approach has not been extended to VARs as far as I know.

15. In contrast to the assumption of equal priors, some lag orders seem less plausible than others to PDQ, but he suspects that may be because he is already familiar with the data. The equal-prior assumption avoids possible bias from a data-based prior.

Pierre Perron (1989) about the effect of the 1970s oil price shock on GDP growth, and he has recently read the working-paper version of Perron and Tatsuma Wada (2009), which updates the argument that there was a change in real GDP's trend growth rate at the beginning of 1973. PDQ also recalls the "Great Moderation," the apparently increased stability of the U.S. economy starting in the early to mid 1980s (e.g., Kim and Nelson 1999, Stock and Watson 2002). PDQ is concerned that the implied heteroskedasticity in real GDP will bias forecast standard errors, and he wonders if there was also change in the growth rate or in the short-run dynamic parameters at that point.

PDQ examines the stability issue in two ways. First, he applies a breakpoint test to all the ARIMA models and to the 15 highest AIC-weighted and 10 highest BIC-weighted VAR models. (The test, which has to be bootstrapped, is too time-intensive to apply to all 625 VARs; the included VARs cover 74% of the AIC weight and 96% of the BIC weight.) Second, PDQ examines some key parameter estimates for a range of different estimation periods. If the estimates change a lot, forecasts based on the longer estimation periods will likely be unreliable.[16]

The breakpoint test is adapted from a quasi-likelihood ratio test discussed by James Stock and Mark Watson (2007, 567-570). The date of the breakpoint is assumed unknown, and the test examines all possibilities within the middle 70% of the dates. If the test rejects homogeneity, it also provides an estimate of the breakpoint date. The key parameter estimates that PDQ examines are the trend rate of real GDP growth and the infinite-horizon impulse responses of real GDP to various shocks. These are not individual parameter estimates but functions of individual parameter estimates that are of particular relevance for forecasting. For example, in the ARIMA model, trend growth is $a \,/\, (1 - \Sigma\varphi_i)$ and the infinite-horizon impulse response to a shock is $(1 - \Sigma\theta_i) \,/\, (1 - \Sigma\varphi_i)$. For both the ARIMAs and the VARs, PDQ computes the trend and impulse response estimates in a recursive manner. He starts with a short estimation period, 1994:1–2008:4, and then moves the start date backwards quarter by quarter until he reaches the longest period of 1948:2–2008:4. If the models are stable, the estimated trends and impulse responses should not change too much in terms of economic importance.

Details of the procedures and outcomes are given in the Appendix. Here I just summarize PDQ's findings and conclusion. For the ARIMAs, the breakpoint tests do not reject the no-break null hypothesis, but trend growth gets substantially higher as the estimation starting point goes back in time, particularly as it moves back into the 1960s. In this way does Perron's trend growth change manifest itself.

16. All procedures in the paper, except one noted in the Appendix, were coded and computed in TSP 5.1. The Appendix contains a link to a web page with the data, TSP files, and Excel files used to get the results in this paper.

In addition, the impulse response values show a great deal of instability in economic magnitude as the estimation start date moves backwards through the early 1980s. Turning to the VARs, the no-breakpoint hypothesis is strongly rejected, with two break dates emerging as most likely: 1973:2 and 1986:3. Trend growth shows the same pattern as in the ARIMAs, and the infinite-horizon impulse responses of y to Δy shocks and to un shocks change suddenly and substantially as the estimation starting date moves to points before 1985.[17] Based on the various results, PDQ concludes that the estimation period for his forecasting equations should start in 1986:3. It's a shame that earlier data containing additional recessions cannot be used, but forecasts from longer data sets are likely to be biased.

Real GDP and the unemployment rate: Unit root or stationary processes?

The literature seems to lean in favor of a unit root in real GDP (e.g., Murray and Nelson 2000, Shelley and Wallace 2011). Regarding the unemployment rate, on first pass it seems that it would be stationary, being bounded by zero and 100 percent, and in theory tending to return to the natural rate. But the natural rate may change and the bounds are not very restrictive. Moreover, the empirical evidence is not clear. For example, Mehmet Caner and Bruce Hansen (2001) strongly reject a unit root in the unemployment rate but conclude there are two "regimes," which consist of two different autoregressive processes that alternate irregularly over time. This is discussed in more detail in the Appendix, as are PDQ's own unit root tests on real GDP and the unemployment rate. PDQ's decision is to stay with Δy and un in the ARIMAs and VARs.

Forecasts

Preliminaries: Linear trends, other forecasts, and post-2008 real GDP values

Mankiw thought that the CEA forecasts looked as if the CEA expected a return to a deterministic linear trend. PDQ therefore fits a linear trend over

17. Because the simultaneity in his structural VAR equations (5) and (6) indicates that structural shocks will not be identifiable, PDQ uses the generalized impulse response procedure of Pesaran and Shin (1998). The procedure applies to the two reduced form VAR equations (with constants omitted) two pairs of reduced form shocks that reflect the correlation in the estimated residuals. See the Appendix.

the 1986:3–2008:4 period to see how the various forecasts relate to it.[18] To help interpret relative movements in the upcoming graphs, Figures 1 to 3, PDQ normalizes all series using the extrapolation of the 1986-2008 trend. Therefore, in the graphs all values are log differences from the extrapolated 1986-2008 trend line. For a second benchmark, PDQ includes a trend of the same slope starting from the average real GDP value for 2007, the year of the most recent peak in the business cycle. In addition to PDQ's forecasts, the graphs include four other forecasts, which are annual: the CEA forecast, the January and February CBO forecasts, and the February Blue Chip forecast.

The graphs also show the actual post-2008 real GDP values from the July 2012 BEA release. PDQ can't know these, of course. They are there for us to see how good the forecasts have turned out to be so far. However, acquiring appropriate values for the comparison is not as straightforward as one might imagine. Later in 2009 and in the two following years, the BEA substantially revised the real GDP series. The revisions changed the base year, but much more important, they substantially increased the reported decline of real GDP in 2008 compared with 2007. How one links the currently reported data with the data used to make the forecasts significantly affects how the actual post-2008 values compare with the forecasted values. I use two approaches.

In the first approach, the quarterly log differences starting in 2009:1 of data reported in February 2012 are added one by one to the log of 2008:4 real GDP to create the post-estimation-period realized quarterly log values. This approach does not penalize the forecasters for not knowing that 2008's quarterly real GDP growth rates would all subsequently be revised downward. But this approach does not correctly depict the negative *annual* growth rate from 2008 to 2009. My second approach does do so by applying the annual 2008-2009 growth rates reported in 2012 to the annual real GDP log value for 2008. The second approach generates annual values. These are the values that probably would be used by the bloggers to settle the bet, had it been accepted.

PDQ's forecasts

PDQ is ready to compute his own forecasts of real GDP. The estimation period is 1986:3–2008:4 and the forecasts are dynamic. The AIC- and BIC-weighted ARIMA forecasts are in Figure 1. The AIC- and BIC-weighted VAR forecasts are in Figure 2.[19] Points labeled "AIC trend value" and "BIC trend value"

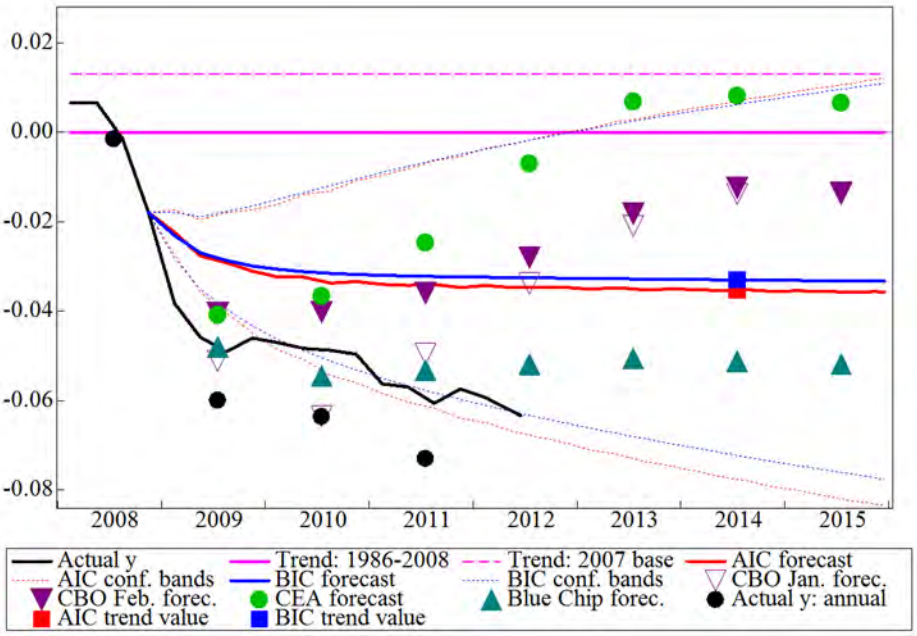
18. Specifically, he estimates an AR(3) model (as chosen by the AIC and BIC) with a constant and a linear trend.

19. A very small number of the VAR models were non-stationary (explosive) and were not used.

for the year 2014 are also plotted; these will be discussed later. The graphs also include AIC- and BIC-weighted plus-minus one standard error confidence bands for PDQ's quarterly forecasts.[20] For more precise assessments, Table 1 gives the numerical values for some of the points in the graphs.

PDQ first notes that by 2013 the CEA forecast exceeds the extrapolation of his 1986-2008 linear trend, and it comes close in 2014 to the higher trend line extrapolated from 2007. This supports Mankiw's statement that the CEA forecast was, in effect, assuming a return to a deterministic linear trend.

Figure 1. ARIMA forecasts, other forecasts, and actual values

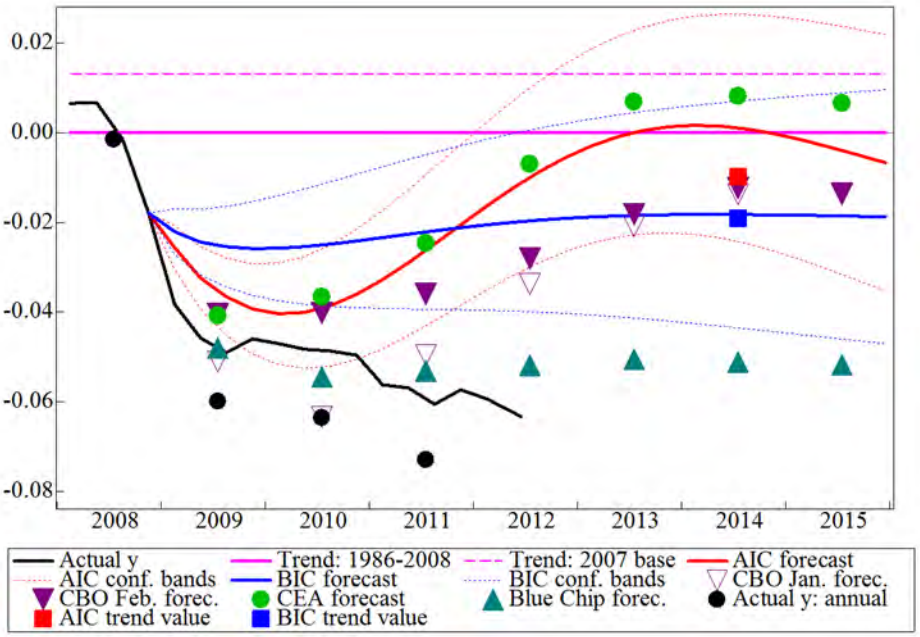


PDQ considers whether he thinks Mankiw would win his bet. None of PDQ's weighted ARIMA or VAR forecasts meet or exceed the CEA forecast for 2013. In fact, although this isn't shown in the graph, not a single one of PDQ's individual ARIMA or VAR forecasts meets or exceeds the CEA forecast. However, the width of the weighted confidence bands suggests to PDQ that a Mankiw victory is not certain. Using the forecast standard error values and assuming normality

20. BIC averaging of forecast standard errors is discussed by Koop and Potter (2003). The accuracy of the confidence bands and the ability to make probability statements using them depends on the estimates being based on serially independent, homoskedastic, and normal residuals. PDQ conducts tests of these assumptions. PDQ finds only a few problems, which he deems of relatively little importance. Details are in the Appendix.

(see the Appendix), the probability that Mankiw loses the bet is 13% for the two ARIMAs and the VAR with BIC weighting and 37% for the VAR with AIC weighting.

Figure 2. VAR forecasts, other forecasts, and actual values



The VAR forecasts, particularly with AIC weighting, are closer to the CEA forecast than are the ARIMA forecasts. PDQ wonders which to believe. He suspects that DeLong and Krugman would favor the VAR results, because the VARs directly address their point that the high unemployment rate of 2008 should be taken into account in the forecasts. But given his own *a priori* uncertainty about the better approach, PDQ would prefer to settle this based on the data and some resulting AIC and BIC weights. But he is not aware of anyone discussing how to get such weights for combining single-equation ARIMA forecasts with two-equation VAR forecasts.[21] Thus he improvises an approach that uses the VAR Δy equations alone to get weights for the VAR forecasts in order to combine them with the single-equation ARIMA forecasts. The procedure is described in the Appendix and combines the 16 ARIMAs with 21 of the original 625 VARs. Since the fundamental issue is whether *un* matters for forecasting, and because PDQ is quite uncertain about this, he gives equal prior probabilities to the set of models with *un* and to the

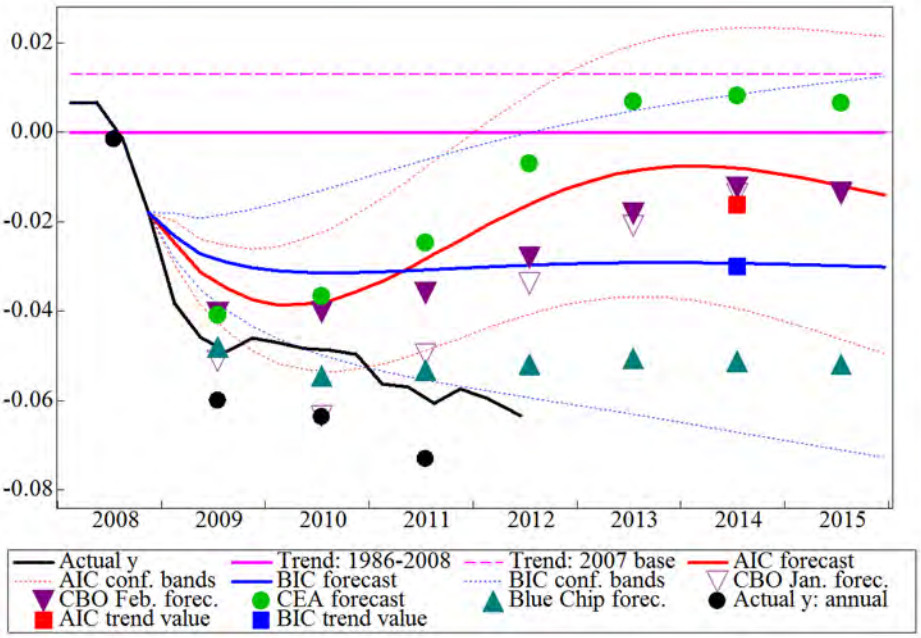
21. This is based on an email from Bruce Hansen dated January 23, 2012, to PDQ's alter ego, the author. Such a technique would have been no more likely to exist three years earlier when PDQ was working.

set without *un* (the "without *un*" models being the ARIMA models and one of the VARs).[22]

The results are in Figure 3. Unsurprisingly, the forecasts lie in between those of Figures 1 and 2, lying 1.6 and 3.6% below the CEA's for 2013. The probability that Mankiw loses his bet according to the overall forecast can also be computed: combined AIC, 28%; combined BIC, 14%.

A key observation is that neither the VARs nor the ARIMAs clearly dominate in the weighting. The AIC weights favor the VAR models while the BIC weights favor the ARIMAs, but these "preferences" are not overwhelming. Thus, the weights do not support DeLong and Krugman's notion that unemployment rates would be necessarily useful in forecasting real GDP.

Figure 3. Combined ARIMA/VAR forecasts, other forecasts, and actual values



22. The 17 equations without *un* (the 16 ARIMAs plus the AR(4) Δ*y* equation) thus each get slightly more weight each (18.5/17) than each of the 20 equations with *un* (18.5/20).

TABLE 1. Various forecasts relative to the extrapolated 1986-2008 linear trend

	2013	2014
Admin (CEA) forecast (Figs. 1-3)	0.007	0.008
Linear trend: 2007 base (Figs. 1-3)	0.013	0.013
Linear trend: 1986-2008 (Figs. 1-3)	0.000	0.000
Linear trend: 2008 base (Figs. 1-3)		−0.002
Linear trend: 2008:4 base (Figs. 1-3)		−0.018
ARIMA AIC forecast (Fig. 1)	−0.035	
ARIMA BIC forecast (Fig. 1)	−0.033	
ARIMA AIC trend (Fig. 1)		−0.035
ARIMA BIC trend (Fig. 1)		−0.033
VAR AIC forecast (Fig. 2)	−0.001	
VAR BIC forecast (Fig. 2)	−0.019	
VAR AIC trend (Fig. 2)		−0.010
VAR BIC trend (Fig. 2)		−0.019
Combined AIC forecast (Fig. 3)	−0.009	
Combined BIC forecast (Fig. 3)	−0.029	
Combined AIC trend (Fig. 3)		−0.016
Combined BIC trend (Fig. 3)		−0.030

Any rebound at all?

PDQ's point forecasts are that real GDP will not recover to the levels forecasted by the CEA for Mankiw's bet year of 2013, but this does not necessarily tell us about rebounds because the forecasted adjustments are not yet complete. To judge rebounds, we need to define a starting point and then compute whether the forecasted long-run equilibrium is higher than that, adjusted for the forecasted trend growth. For example, some of PDQ's forecasts for the long-run recovery are actually more pessimistic than his 2013 values because in those cases 2013 is near the peak of an oscillation in the recovery path.

What starting point should PDQ apply? The CEA's February discussion was all in terms of annual growth rates, and 2008 was the base year for the CEA's first annual growth forecast, so 2008's annual real GDP would certainly seem reasonable as the rebound starting point. A variation would be to use the fourth quarter of 2008. Because this value is less than the annual value, it is a less stringent criterion for declaring a rebound. Finally, one could decide that if real GDP is forecasted to eventually recover at least somewhat from whatever low is ultimately reached after 2008, then a rebound is forecasted. Then all we need is some amount of forecasted

recovery from the troughs that all the forecasts show in 2009 or 2010. But the CEA was forecasting that a rebound would occur relative to 2008, so PDQ sticks to that.

PDQ then computes forecasted long-run trend values using the AIC and BIC weights. The values are plotted for the year 2014 in the Figures 1 to 3 to allow easy comparison with the CEA forecasted long-run trend value, which the annual CEA forecast reaches the same year. The long-run values can then be graphically compared with the two 2008 starting values by imagining a horizontal line drawn from the long-run value in 2014 back to the values in 2008. Or they can be compared numerically by applying a few computations to results in Table 1.

These comparisons assume that the various models' forecasted long-run growth rates are the same as the linear-trend annual growth rate (2.728%) used to normalize the values in the graphs. PDQ's ARIMA and VAR growth rates are not exactly equal to that, but they are close enough (2.697% to 2.736%) that applying the model-specific growth rates makes no discernible difference to the computed rebound sizes forecasted by PDQ's models. These growth rates are essentially the same as the Blue Chip long-run growth rate of 2.7%.

PDQ finds no rebounds relative to the annual 2008 value in the weighted forecasts of *any* of his models. The only weighted forecast rebounds he finds are relative to the 2008:4 starting point using the VAR and the combined ARIMA/VAR forecasts with AIC weighting. The rebounds are rather small, amounting to 0.8% (VAR) or 0.2% (combined ARIMA and VAR).[23] Note that real GDP's move to the high point of the AIC-weighted VAR forecast path near the end of 2013 is not a rebound in PDQ's view because it is temporary.

In contrast to PDQ's small to nonexistent forecasted rebounds, the CEA forecasted rebound is 2.6%, net of the same 2.728% long-run growth rate used for the other rebound computations. And if the CEA's own forecasted long-run growth rate of 2.6% is used instead for computing the CEA net forecasted rebound, it becomes 3.5%. To put it another way, the CEA forecast for 2013 rebounds almost exactly to a trend line extrapolated from 2007's real GDP using the CEA's growth rate of 2.6% (again consistent with Mankiw's trend reversion point).

Tyler Cowen has recently asked, "When was it obvious our recovery would be so slow?" (Cowen 2012). This was in response to several recent discussions involving the idea that rebounds tend to be stronger after deep recessions than after shallow ones (reminiscent of the 2009 CEA statement). PDQ's prediction in early 2009 of little to no rebound suggests an answer to Cowen's question, but the recent

23. This does not mean that PDQ's models predict stronger rebounds can never happen. Some impulse response analysis available from the author upon request shows the possibility of larger partial rebounds than in the present case.

discussion more specifically regards recovery from the now-known official ending point of the recession, June 2009.[24] We therefore should look at PDQ's forecasted recoveries from his post-2008 predicted troughs. His BIC-weighted forecasts show little to no recovery. Thus, in early 2009 it would have been predicted by (if not completely obvious to) a believer in PDQ's BIC-weighted results that the recovery would be very slow. A believer in PDQ's AIC-weighted results would have found slow recovery less obvious. Nevertheless, the AIC believer would not have predicted a recovery strong enough to get back to the 1986-2008 trend line.

How are all the forecasts doing?

I now turn to something 2009's PDQ cannot know: how all the forecasts are doing so far. The Blue Chip forecast was the most pessimistic and is the most accurate through 2011, marginally beating the January CBO forecast using the mean squared error criterion (MSE = 0.000207 versus 0.000209). The February CBO forecast, revised to account for the stimulus, does worse (MSE = 0.0026). The CEA and all PDQ forecasts but one have MSEs in the range of 0.0034 to 0.0039, with PDQ's VAR-BIC trailing at 0.0053. Basically, PDQ's forecasts are the least successful of the forecasts at catching the depth of the recession in 2009.[25]

Looking beyond 2011, if real GDP does not soon experience what would be a remarkable burst of growth, then the relative performance of PDQ's ARIMA forecasts will rise, and PDQ's VAR forecasts will prove inferior (and DeLong and Krugman's insistence that the unemployment rate is essential to the forecasting will, again, be unsupported). Finally, the CBO's forecast of a much-lower growth rate of 2.2% later in the decade (given in OMB 2009) is starting to look very good.

24. Cowen cites Cochrane (2012b), who cites Taylor (2012), who cites Bordo and Haubrich (2012). Cochrane (2012b) draws various linear trends for real GDP (see also Cochrane 2012a) and thus apparently believes in linear trend stationarity and the implied full recoveries doubted by Mankiw (2009b). Cochrane (2012b) also points out that each year since 2009 the Administration forecasts have been more optimistic than the Blue Chip consensus forecasts. Bordo and Haubrich (2012) analyze U.S. recessions going back to the 1880s and find the pattern of stronger recoveries after deeper recession. But, as they mention, they do not address whether the recoveries are ever strong enough to get back to any former trend.

25. It's not clear whether the February Blue Chip forecast for 2009 and 2010 accounted for the Obama stimulus. It was, however, more *pessimistic* than the Blue Chip January forecast. The Blue Chip accuracy so far does not prove that the Blue Chip consensus would always be more accurate than other forecasts. To do so would require a study of its own.

Final remarks

In 2009, using some fairly standard techniques with a few refinements from the recent literature, my hypothetical time series econometrician, PDQ, confirms Mankiw's skepticism about the CEA's early 2009 optimistic forecasts of real GDP. Mankiw would likely win his bet, with a probability of 72% to 86% according to PDQ's overall estimates. In this instance, at least, Mankiw's intuition appears to PDQ to be superior to DeLong and Krugman's. Perhaps Mankiw's doubts were driven not only by his unit root point but also by his well-known skepticism regarding the Obama stimulus plan (e.g., Mankiw 2008, 2009a). And, although they didn't provide probabilities, the CBO and Blue Chip consensus forecasts, readily available at the time, also suggested Mankiw would win. It is therefore quite unclear to PDQ how DeLong and Krugman could have been so dismissive of Mankiw's skepticism and so certain of the CEA's forecasted rebound.

Appendix

Data and code for replication

Data and computer code to aid those interested in replication are available via econjwatch.org (**link**). To run all code as is, the reader would need Excel, TSP 5.1 (somewhat earlier versions should also work), and R. The reader can check the CBO data interpolations and recreate the breakpoint tests, weighted forecasts, all graphs, heteroskedasticity tests, normality tests, and unit root tests.

Mankiw's other two points

In addition to his primary point about unit roots and the permanency of shocks, Mankiw (2009b) raised two other points about the CEA (2009) rebound discussion. The CEA presented data for and a graph of the regression of the real GDP "rebound" (measured as the subsequent two-year growth rate) on the percentage peak-to-trough real GDP decline for eight U.S. post-war recessions. The regression line shows higher rebounds associated with deeper troughs, with a *t*-statistic of 1.97.

Mankiw's (2009b) first non-unit-root point involved the interplay between a sample selection issue and the possibility of heteroskedasticity. His write-up was somewhat ambiguous, but he kindly clarified matters for me by email. As in his blog, let G = real GDP growth and V its variance. In essence, the CEA regressed G(t) on G($t - 1$). But the CEA only used time periods consisting of recession followed by recovery (negative G($t - 1$) followed by positive G(t)). Suppose there is positive ARCH(1).[26] Then big V($t - 1$) means that somewhat big V(t) is likely. Furthermore, an observed G($t - 1$) that is far from its mean is more likely to come from a distribution with big V($t - 1$) than otherwise. Thus, the more negative is G($t - 1$), the more positive G(t) is likely to be, biasing the CEA regression in the direction of showing big recessions followed by big recoveries.

Mankiw's second non-unit-root point was that the CEA regression data omitted the 1980 recession, but he did not elaborate much on the problem. I do so here. The CEA regression also omitted the 1949 recession. The CEA (2009) gave the following rationale for omitting these data points:

26. The assumption that the heteroskedasticity was ARCH was not explicit in the blog.

> The 1949 recession is excluded because it was followed by the outbreak of the Korean War, resulting in exceptionally rapid growth. The 1980 recession is excluded because it was followed by another recession, resulting in unusually low growth.

In the case of the 1980 omission noted by Mankiw, it would seem that the CEA dropped the data point *because* it clearly contradicted the CEA's hypothesis. But the 1949 omission involves the same problem, because not only was that recession followed by strong growth, it was a shallow recession. If one reruns the regression with either or both of these two recessions included, the statistical significance of the rebound is completely eliminated (*t*-values fall to 1.06 or less). One might add that a regression based on only eight observations is not very believable.

Lead-up to and background for the DeLong email

In a September 26, 2011, *Wall Street Journal* essay, Harold Cole and Lee Ohanian wrote:

> The Federal Reserve Board's Index of Industrial [P]roduction rose nearly 50% between the Depression's trough of July 1932 and June 1933. This was a period of significant deflation. Inflation began after June 1933, following the demise of the gold standard. Despite higher aggregate demand, industrial production was roughly flat over the following year.

A day later, Krugman (2011) posted a graph of industrial production and the producer price index (PPI) over 1929-1936 that showed generally positive co-movements including a net rise in both from July 1932 to June 1933. He wrote:

> You might think that this looks pretty straightforward: output shrank when prices were falling, grew when they were rising, which is what a demand-side story would predict. But Cole and Ohanian focus on the month-to-month wiggles in 1932-33—conveniently omitting wiggles that went in an inconvenient direction—to claim that demand had nothing to do with it. This goes beyond holding views I disagree with (as does much of what happens in this debate). This is a deliberate attempt to fool readers, demonstrating that there is no good faith here.

On his blog, Stephen Williamson (2011) reacted:

> Hal and Lee are two thoughtful and careful economists. I don't agree with everything they have ever said, but to call them liars is appalling.

DeLong (2011) joined in:

> Williamson should be much more unhappy at Cole and Ohanian's claim that July 1932-June 1933 was "a period of significant deflation." The PPI in July 1932 is 11.1. The PPI in June 1933 is 11.2. Cole and Ohanian may be the only people who have ever managed to call a period during which the price level rose as "one of significant deflation".

On September 28, after reading these exchanges, I wondered how Cole and Ohanian could make such a mistake (if not attempting to "fool readers") and looked up the price data for myself. I found that, while the overall PPI did indeed have a net rise over the period as DeLong pointed out (about 2%), the CPI had a net fall of about 7%. I presumed that Cole and Ohanian based their "deflation" claim on the CPI rather than on the PPI.[27] My attempted comment on DeLong's blog was something to the effect that it was misleading to mention only that the PPI had risen a bit and so DeLong ought to have also mentioned that the CPI had fallen a lot. Although it was certainly a critical comment, I do not think I was obnoxious in expressing it, but since he did not publish my comment and I had not thought to keep a copy of what I filled into the comment form, I cannot prove it. Anyway, his emailed response was, "It's not a period of 'significant deflation' if one of your two price indexes is not falling. Shame on you for trying to confuse the issue." "Shame on me?" I wrote back. "I wouldn't have thought bringing a bit more information to bear was confusing the issue."

The revised CBO forecasts

The revised CBO forecasts (Elmendorf 2009) were given as percentage revisions to the level of real GDP in the fourth quarters of 2009-2019. The CEA, January CBO, and Blue Chip forecasts, and the Mankiw bet, were, however, all in terms of calendar years. To compare its own forecasts with the revised CBO forecasts, the CEA said that it "interpolated the impacts by quarter" (CEA 2009, footnote 3) to convert to calendar year terms, but the CEA did not precisely describe the procedure, and the CEA only applied it to 2009 and 2010. Therefore, I conducted my own conversion.

27. To recreate what would have been found in September, 2011, when the blogs and my attempted comment were written, I used the September 2011 vintage data from ALFRED.

Stimulus spending began in early March, and I assume that the initial date of its effect on GDP and thus the CBO's revision is one month later.[28] I assume that calendar-year real GDP values are centered at the beginning of July and fourth-quarter values in the middle of November. I divide the year into 36 periods where the 19th is deemed to be the beginning of July and the 32nd the middle of November. I then place the January CBO calendar-year values in period 19 and interpolate to get the remaining 35 values (except that actual third-quarter values and the CBO's fourth-quarter estimates as of January 2009 are used for the second half of 2008). In the interpolations, I also impose some smoothing of the January CBO growth rates from year to year. Next, I place in each year's 32nd period the midpoint of the low and high percentage increases to fourth-quarter levels in the CBO's February revision, and I interpolate to get the remaining 35 percentage increases. The resulting 36 percentage increases for each year are applied to the interpolated January CBO values, and the values for each year's 19th period are used to get the February CBO revision of annual real GDP growth rates on a calendar-year basis. My procedure gives growth rates for 2009 and 2010 that are similar to the CEA's estimates, and it also generates growth rates for the remaining years.[29]

AIC and BIC weighting

Weighting a set of single equations is described in Hansen (2007). The formulas here are generalized to allow for weighting multi-equation models and to allow for unequal priors. Define the AIC criterion for model m with k constants and slopes and a sample size of n as $\mathrm{AIC}_m = n \ln\left|\hat{\Sigma}_m\right| + 2k$ where $\hat{\Sigma}_m$ is the estimated error variance (single equation) or covariance matrix (multi-equation). Define the BIC criterion for model m with k constants and slopes as $\mathrm{BIC}_m = n \ln\left|\hat{\Sigma}_m\right| + \ln(n)k$. Then, letting IC be in turn either AIC or BIC, and pr the prior probability, the weight for model m in a set of M models is

$$w_m = pr_m \cdot \exp(-\tfrac{1}{2}\mathrm{IC}_m) \Big/ \sum_{j=1}^{M} pr_j \cdot \exp(-\tfrac{1}{2}\mathrm{IC}_j).$$

28. See the "Stimulus Speed Chart" from ProPublica (Larson, Flavelle, and Knutson 2010).
29. The reader may wonder why I don't just use the CEA version of the CBO revision for 2009 and 2010. I do not because I compute that the 2009-2010 CEA growth rate for the revised CBO values is slightly too high to be consistent with Elmendorf (2009).

In computation, if the information criteria values are sufficiently large in absolute value, as they turn out to be in PDQ's application, they need to be normalized by subtracting some typical value from each one before applying the formula, or numerical errors occur.

Breakpoint tests

Stock and Watson (2007) present the QLR test as based on an F test. One computes F statistics for the null that all constant and slope parameters are constant against the alternative that at least one changes for all break dates in the middle 70% of the overall estimation period. The largest resulting F value is then compared to a nonstandard sampling distribution. Instead of an F test, PDQ computes a likelihood ratio test (for the ARIMAs) and a Wald test (for the VARs). Likelihood ratio and Wald tests are asymptotically equivalent (Greene 2003, 484), but in the case of the VARs, PDQ wants to test stability for a number of subsets of parameters because the overall null of stability was rejected. It is easier to do this with Wald tests because the model only needs to be estimated once instead of multiple times as with the likelihood ratio version. However, the subset results are not very revealing and are not reported here.

Because of the almost certain presence of heteroskedasticity in the post-War data, PDQ applies a recursive wild bootstrap approach to get p-values instead of comparing the likelihood ratio and Wald statistics with tabled distributions. Silvia Gonçalves and Lutz Kilian (2004) discuss the validity of the recursive wild bootstrap for autoregressive models. Assume autoregressive order r and let $Y_{t-1} = (y_{t-1}, \ldots, y_{t-r})'$. Estimate regression model $y_t = Y_{t-1}'\hat{\varphi} + \hat{\varepsilon}_t$ (with constants and trends included as desired). The estimated coefficients and residuals comprise the data generating process (DGP), with which one builds up simulated data sets using $y_t^* = Y_{t-1}^{*\prime}\hat{\varphi} + \hat{\varepsilon}_t\eta_t$ where η_t is i.i.d.(0,1). PDQ follows Herman Bierens's EasyReg econometrics program procedure by generating η_t as a standard normal variable.[30] One then applies the desired statistical test to the simulated data sets to generate the sampling distribution. PDQ extends the procedure to the VARs and ARIMAs. The lag orders of the DGP are thus the same as in the model being tested, and the residuals are normal with a heteroskedastic component measured by $\hat{\varepsilon}_t$. Also following Bierens, PDQ uses the first few actual data values for the initial lag values in the recursive process.

30. The current version is Bierens (2011), but the wild bootstrap procedure has been in the program since well before PDQ's working period in 2009.

For the two-equation VARs, PDQ needs a pair of residuals for each period t in the place of the single $\hat{\varepsilon}_t \eta_t$ in the univariate case. The pair needs to reflect the heteroskedasticity in each equation *and* the covariance of the errors between equations. PDQ first computes moving three-period covariance matrices using the estimated residuals and then randomly selects a pair of values from the bivariate normal distribution with covariance matrix centered on period t.

Thus far, the creation of bootstrapped data sets has been described. PDQ generates 5,000 of them for each of the 16 ARIMAs and each of the 15 AIC-weighted and 10 BIC-weighted VARs (two of these are the same, so there are 23 tested VARs in all). For each model, then, PDQ has an actual breakpoint statistic for each break date and 5,000 simulated breakpoint statistics for each date. He uses the actual value and set of simulated values for each date to compute a bootstrapped p-value for that date. The date of the most significant p-value is the Stock-Watson estimate of the break date, if there is one. But to determine if there is one, an overall level of significance must be generated. Suppose the lowest p-value across all the dates is 0.03. If the null of stability is true, it is much more likely than 0.03 to get a p-value of this seemingly low value because the test has been run 167 times to cover the middle 70% of dates. Thus, we want an overall p-value that expresses how unlikely 0.03 really is. Now, each breakpoint has 5,000 simulated test statistics that can be converted into 5,000 p-values between 0 and 1. As a result, each of the 5,000 replications has a set of 167 p-values. PDQ's overall p-value is the fraction of replications with at least one p-value less than or equal to 0.03.[31]

As noted in the main text, none of the ARIMA tests were close to significance and so they are not reported here. Tables 2 and 3 give the VAR breakpoint test results for the highest weighted VAR models. Figures 4 and 5 graph the individual breakpoint p-values for the highest weighted AIC and BIC models in order to show how the results seem to suggest two breakpoints.

The main text mentions PDQ's alternative evidence of structural instability in the form of recursive estimates of trends and infinite horizon impulse responses. Figures 6 and 7 give the AIC- and BIC-weighted ones for the ARIMAs. The shocks for impulse responses are contractionary and the responses are cumulated and thus for y, not Δy. The dates on the horizontal axis refer to the beginning of the estimation period. Figures 8 to 10 give the AIC- and BIC-weighted trends and Pesaran-Shin generalized impulse responses in y for the VARs. (Because the *un* equations in the VARs are stable, the infinite horizon responses in *un* to shocks are zero.)

31. In this, PDQ is applying a procedure not specifically seen in the literature as of 2009 (so far as I know), so in this respect he seems to be violating my rule that he use only techniques available then. I therefore posit a creative burst on his part. Or perhaps he saw Cushman (2008) or a working paper version of Cushman and Michael (2011), in which similar approaches are developed.

In Figure 10, the ultimate response of *y* to a generalized shock in *un* for the estimation period commencing 1986:3 is negative. This means that the point estimate for the effect on *y* is for an incomplete rebound if any. However, this does not really tell us about the response to a transitory shock, because it is not possible to identify the transitory shock.

TABLE 2. Breakpoint test results for the top AIC-weighted VAR models

Lag orders	Model weight	Overall p-value	Break point	Break point 2
2, 3, 3, 3	0.171	0.019	86:3	75:4
2, 3, 2, 3	0.075	0.043	86:3	75:4
2, 3, 3, 4	0.057	0.000	86:3	75:4
2, 4, 3, 4	0.055	0.024	86:3–86:4	75:4
3, 3, 2, 3	0.049	0.017	86:3	75:4
0, 4, 3, 4	0.046	0.004	86:2–86:3	73:3, 75:4
1, 3, 3, 3	0.045	0.005	85:2–86:4	75:4
0, 3, 3, 3	0.043	0.007	85:3–87:1	75:4
2, 4, 2, 3	0.037	0.066	86:3	75:4
2, 4, 3, 3	0.034	0.032	86:3	75:4
2, 3, 4, 3	0.032	0.034	86:3	75:4
1, 4, 3, 4	0.030	0.011	86:1–86:3	75:4
3, 3, 3, 3	0.025	0.021	86:3	75:4
0, 4, 2, 3	0.020	0.030	86:3	75:4
1, 3, 2, 3	0.020	0.025	86:3	75:4

Note: The lag order column gives the Δy equation lag orders for Δy and *un*, then the *un* equation lag orders for Δy and *un*.

Figure 4. VAR breakpoint p-values: AIC

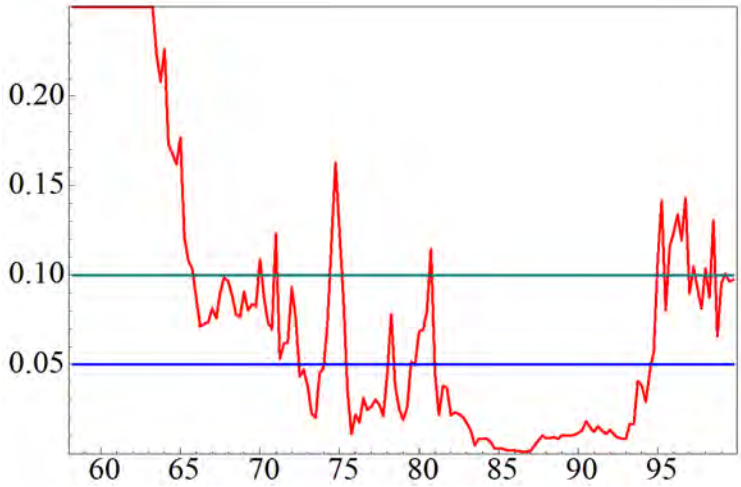


TABLE 3. Breakpoint test results for the top BIC-weighted VAR models

Lag orders	Model weight	Overall p-value	Break point	Break point 2
0, 2, 2, 2	0.377	0.010	73:2–73:3	86:3
0, 3, 2, 3	0.245	0.024	86:3	73:3
0, 2, 1, 2	0.166	0.008	73:2–73:3	86:3
1, 2, 2, 2	0.050	0.027	73:2	86:3
0, 3, 1, 2	0.042	0.018	73:2	86:2
0, 2, 2, 3	0.029	0.008	75:4	86:3
1, 2, 1, 2	0.022	0.053	73:3	86:3
0, 3, 3, 3	0.017	0.007	85:3–87:1	75:4
0, 4, 2, 3	0.008	0.030	86:3	75:4
0, 3, 2, 2	0.008	0.021	73:3	86:3
Note: Lag orders are denoted as in Table 2.				

Figure 5. VAR breakpoint p-values: BIC

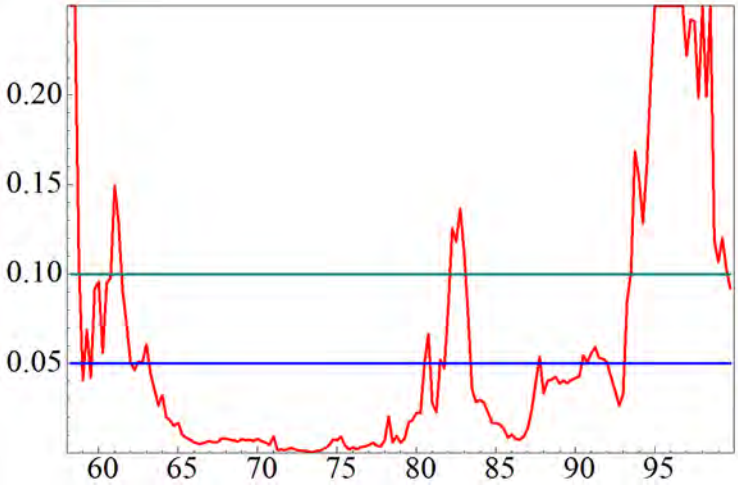


Figure 6. Recursive ARIMA trends

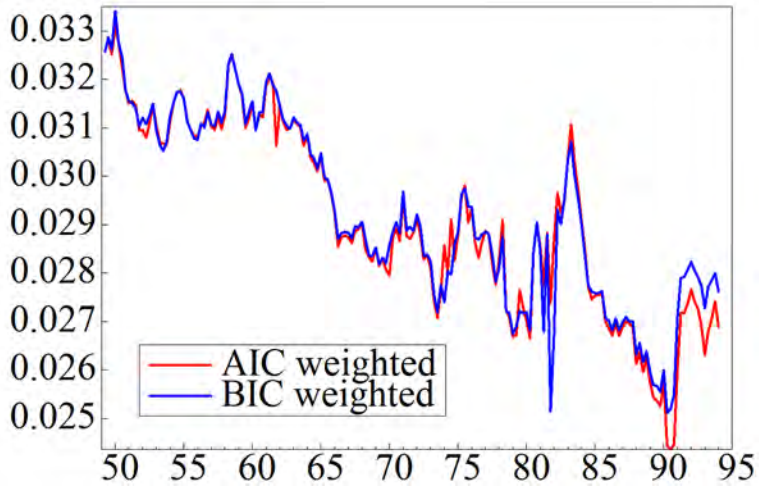


Figure 7. Recursive ARIMA impulse responses

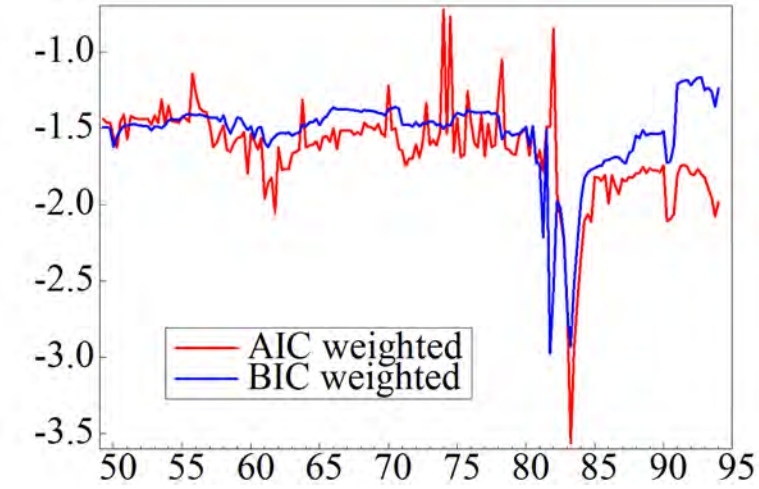


Figure 8. Recursive VAR trends

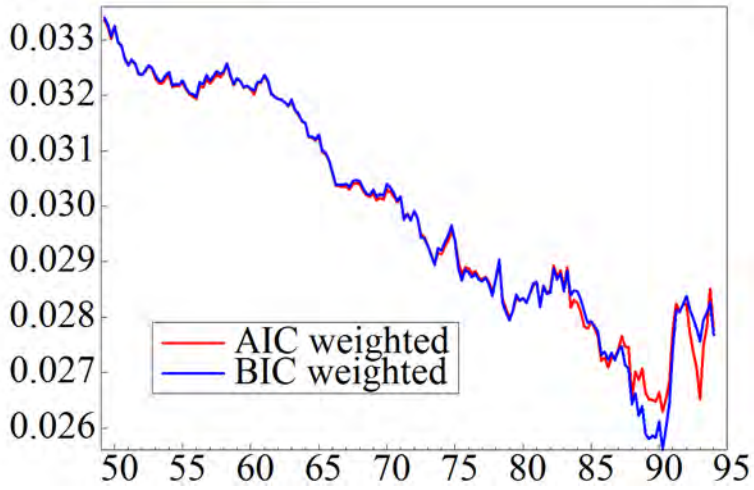


Figure 9. Recursive VAR impulse responses to Δy

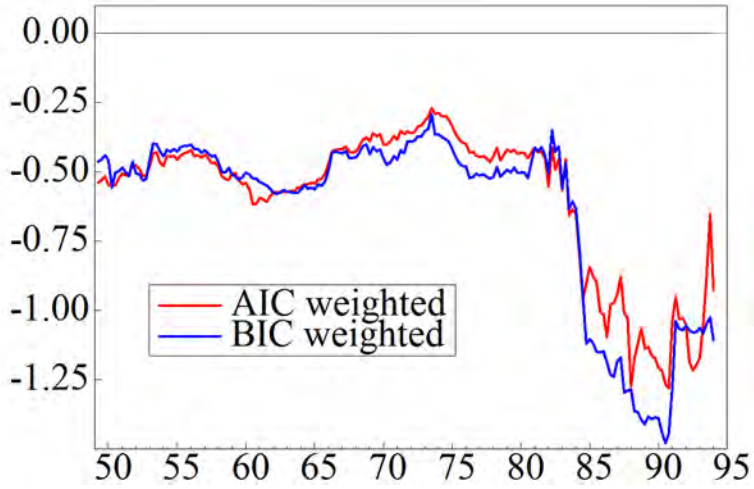
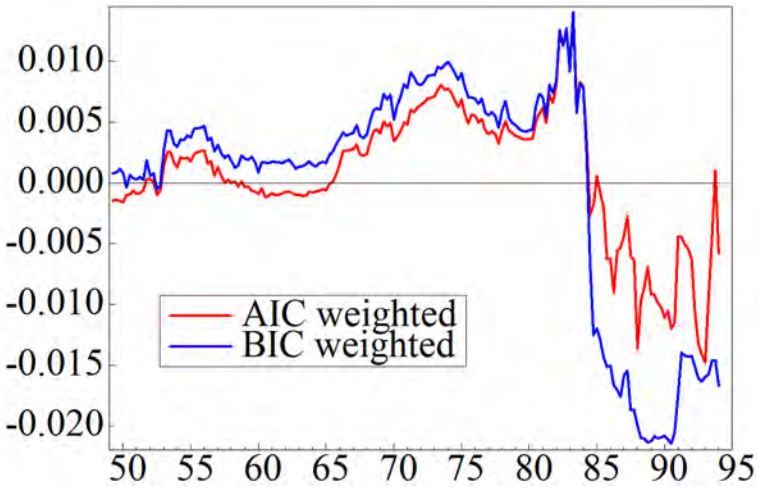


Figure 10. Recursive VAR impulse responses to *un*



The Pesaran-Shin generalized shock responded to here is a pair of reduced form shocks, one for each equation, whose relationship to each other is based on the correlation of the reduced form VAR residuals. For an *un* shock, the shock to the *un* equation is 1.0 and the simultaneous shock to the Δy equation is the slope coefficient in the regression of the Δy equation residuals on the *un* equation residuals. For a Δy shock, the shock to the Δy equation is 1.0 and the simultaneous shock to the *un* equation is the slope coefficient in the regression of the *un* equation residuals on the Δy equation residuals.

Unit root tests

For results specific to his data set, PDQ applies the DF-GLS-MAIC test of the unit root null (Elliott, Rothenberg, and Stock 1996, Ng and Perron 2001) and the modified KPSS test of the mean or trend stationarity null (Harris, Leybourne, and McCabe 2007). The modification of the KPSS test is to filter out a near unit root (under the null) to reduce size distortion. A wild bootstrap (as in the breakpoint tests, but univariate) is used to get p-values. The DGP lag orders are determined by the MAIC lag choice of the DF-GLS-MAIC test. When bootstrapping the DF-GLS-MAIC test, in each replication the test chooses the lag order, which may therefore be different than that of the DGP (leading to "exact" p-values, as in Murray and Nelson 2000). The modified KPSS test uses a sample-size determined lag. PDQ uses $lag = \text{int}(12(T/100)^{0.25})$, a typical choice. To filter out the presumed near unit root, he uses a rho value of 0.90.

For *y*, the unit root null is not rejected (p-value = 0.795), confirming the Murray-Nelson (2000) finding. However, the stationarity null for *y* is *also* not rejected (around a linear trend, p-value = 0.366). For Δy the mean stationarity null is not rejected (p-value = 0.490) and the unit-root no-drift null *is* rejected (p-value = 0.017). Probably, then, Δy meets the assumptions of PDQ's ARIMAs and VARs. However, if *y* itself is actually trend stationary, then both the ARIMA and VAR models above would contain a moving average unit root ($\theta_1 = 1$, $c_1 = 0$, $\psi_{t-1} = 0$ in equations (3), (7), and (8)). The ARIMA procedure PDQ is going to use can handle this.[32] The VAR, on the other hand, would need a long lag order to approximate the MA unit root process.

For *un*, mean stationarity is not rejected (p-value = 0.448), but neither is the unit root null (p-value = 0.237). Contrary to Krugman (2009), it is therefore not "very clear" that the unemployment rate has no unit root. For further evidence, PDQ looks at the weighted symmetric tau unit root test (Pantula, Gonzalez-Farias, and Fuller 1994) as implemented in TSP 5.1. The lag choice is from an AIC-plus-two rule. The test generates a weak unit root rejection (p-value = 0.069). Next, in light of the Caner and Hansen (2001) conclusion of two regimes for the autoregressive process (a threshold autoregression, TAR) for the adult male unemployment rate over the 1956-1999 period, PDQ applies their test to his 1986-2008 data (in monthly form), using code on Bruce Hansen's web page.[33] The possible second regime in Caner and Hansen's application, which PDQ follows exactly, is triggered by recent large increases (beyond some threshold) in the unemployment rate. Both the homogeneous-regime and unit-root nulls are rejected with (coincidentally equal) bootstrapped p-values of 0.012. If there are, in fact, two autoregressive regimes, PDQ's VARs may be misspecified. However, the two-regime result may not hold in the two-equation VAR. Furthermore, Caner and Hansen (2001), who were focusing on the econometric procedure, provide no economic rationale as to why the short-run dynamics and mean of unemployment might change when it has been rising a lot recently, regardless of its current level or other factors. Overall, PDQ concludes that assuming a homogeneous *un* process remains reasonable.

32. PDQ uses the exact maximum likelihood procedure of Mélard (1984) in the econometrics program TSP 5.1. This allows $\hat{\theta} = 1$ and gets correct standard errors, unlike the traditional conditional likelihood approach in such a case. Furthermore, a long lag order is not necessary, contrary to the case of a purely autoregressive model.

33. Gauss, Matlab, and R code are provided by Hansen (**link**).

Serial correlation, heteroskedasticity, and normality tests

PDQ applies, where lag orders permit, Q(4), Q(6), and Q(8) Ljung-Box tests for serial correlation to the 16 ARIMA models. He applies AR(4) Wald tests to the 15 highest AIC-weighted VAR models and the 10 highest BIC-weighted models.[34] Q statistics are part of the standard ARIMA output in TSP 5.1. For the VAR models, both univariate and multivariate Wald tests are computed. The two univariate AR(4) tests are constructed by regressing each equation's residuals on its own one-to-four lagged residuals and a constant. The multivariate AR(4) test is the joint Wald test of significance of all non-constant coefficients in the two-equation system of the Δy and *un* equation residuals regressed on four lags of the residuals from both equations (therefore 16 jointly tested coefficients). The only evidence of serial correlation is in two BIC-weighted VARs where the Δy equation has no right-hand-side variables at all (other than a constant). The two VAR models' combined weight is only 15%. Because of the statistical insignificance of most of these results, I don't give them in a table, but they are available.

PDQ applies ARCH(4) tests for heteroskedasticity and Jarque-Bera (JB) tests for normality. The procedure is the same as for the univariate and multivariate AR serial correlation tests, except that the residuals are squared. P-value results are in Tables 4 to 6.

TABLE 4. ARIMA heteroskedasticity and normality test p-values

AR, MA orders	ARCH(4)	JB test
0, 0	0.534	0.133
0, 1	0.793	0.485
0, 2	0.309	0.520
0, 3	0.355	0.525
1, 0	0.770	0.657
1, 1	0.854	0.674
1, 2	0.860	0.709
1, 3	0.671	0.651
2, 0	0.677	0.685
2, 1	0.429	0.660
2, 2	0.592	0.706
2, 3	0.468	0.684
3, 0	0.562	0.668
3, 1	0.496	0.680
3, 2	0.549	0.696
3, 3	0.692	0.479

34. The top-weighted models are different from the ones for the breakpoint tests because the estimation period is different.

The ARIMAs have no significant heteroskedasticity or non-normality. Some of the VARs do show heteroskedasticity, but it appears to be in the *un*, not the Δy equation. Therefore, forecast confidence bands for Δy are likely to be tainted less than otherwise. The VAR normality tests raise no concerns. Moreover, they call into question the heteroskedasticity rejections, because heteroskedasticity will tend to be interpreted by the JB test (which assumes homoskedasticity) as a violation of normality. Similar lag structures explain many of the similarities among the results.

TABLE 5. AIC-weighted VAR heteroskedasticity and normality test p-values

Lag orders	Model weight	ARCH Δy	ARCH *un*	ARCH multi	JB Δy	JB *un*
1, 3, 1, 3	0.122	0.884	0.313	0.156	0.782	0.290
1, 3, 2, 3	0.077	0.884	0.205	0.172	0.782	0.238
0, 3, 1, 3	0.058	0.906	0.356	0.140	0.883	0.254
2, 3, 2, 3	0.049	0.825	0.161	0.085	0.849	0.220
0, 3, 2, 3	0.037	0.906	0.216	0.131	0.883	0.210
3, 3, 2, 3	0.034	0.680	0.161	0.175	0.738	0.220
2, 0, 4, 2	0.026	0.684	0.011	0.000	0.723	0.201
2, 0, 2, 3	0.026	0.684	0.042	0.001	0.723	0.158
1, 3, 1, 4	0.024	0.884	0.234	0.107	0.782	0.289
1, 3, 2, 4	0.024	0.884	0.104	0.093	0.782	0.239
2, 0, 2, 2	0.022	0.684	0.023	0.001	0.723	0.205
2, 3, 1, 3	0.022	0.869	0.313	0.110	0.808	0.290
1, 4, 1, 3	0.021	0.899	0.313	0.171	0.752	0.290
1, 3, 4, 3	0.020	0.884	0.105	0.163	0.782	0.255
1, 4, 2, 3	0.016	0.904	0.200	0.197	0.741	0.236
Note: Lag orders are denoted as in Table 2.						

TABLE 6. BIC-weighted VAR heteroskedasticity and normality test p-values

Lag orders	Model weight	ARCH Δy	ARCH *un*	ARCH multi	JB Δy	JB *un*
1, 0, 1, 2	0.242	0.769	0.011	0.000	0.693	0.238
2, 0, 2, 2	0.122	0.684	0.023	0.001	0.723	0.205
0, 2, 0, 2	0.119	0.992	0.168	0.215	0.928	0.155
0, 0, 0, 2	0.107	0.534	0.039	0.003	0.160	0.072
1, 0, 1, 3	0.073	0.754	0.036	0.000	0.796	0.200
0, 2, 1, 2	0.049	0.992	0.062	0.015	0.928	0.176
0, 0, 1, 2	0.044	0.534	0.007	0.000	0.160	0.104
1, 0, 2, 2	0.040	0.769	0.008	0.000	0.693	0.187
0, 3, 1, 3	0.026	0.906	0.356	0.140	0.883	0.254
2, 0, 1, 2	0.016	0.709	0.017	0.000	0.682	0.231
Note: Lag orders are denoted as in Table 2.						

Overall weighting of the ARIMA and VAR models

PDQ computes the 16 ARIMA AIC and BIC values as before, but for the VAR models he computes AIC and BIC values not for the 625 VARs, but for the 25 distinct Δy AR equations that appear in the VARs. But four of them, the ones with zero to three Δy lags and no *un* lags, are the same as the four ARIMA equations with no moving average terms, so the duplicate AR equations are dropped.[35] PDQ thus gets AIC and BIC weights for 37 equations. The 16 ARIMA forecasts are, of course, the same as before. To get the VAR forecasts, PDQ adds to each Δy equation the *un* equation that gives the highest VAR model probability among the 25 VARs with the given Δy specification.

References

Bierens, Herman J. 2011. EasyReg International. Department of Economics, Pennsylvania State University (State College, Pa.). Link

Bordo, Michael D., and Joseph G. Haubrich. 2012. Deep Recessions, Fast Recoveries, and Financial Crises: Evidence from the American Record. *NBER Working Paper* No. 18194. National Bureau of Economic Research (Cambridge, Mass.).

Box, George, and Gwilym Jenkins. 1970. *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.

Campbell, John Y., and N. Gregory Mankiw. 1987a. Are Output Fluctuations Transitory? *Quarterly Journal of Economics* 102(4): 857-880.

Campbell, John Y., and N. Gregory Mankiw. 1987b. Permanent and Transitory Components in Macroeconomic Fluctuations. *American Economic Review* 77(2): 111-117.

Caner, Mehmet, and Bruce E. Hansen. 2001. Threshold Autoregression with a Unit Root. *Econometrica* 69(6): 1555-1596.

Clark, Todd E., and Michael W. McCracken. 2006. Forecasting with Small Macroeconomic VARs in the Presence of Instabilities. *Finance and Economics Discussion Series* 2007-41. Federal Reserve Board (Washington, D.C.).

Cochrane, John. 2012a. Just How Bad Is the Economy? *The Grumpy Economist*, July 31. Link

35. Because ARIMA estimation treats the lag values for the initial estimation periods differently than OLS estimation does, the overlapping models' estimates are not actually identical, but the differences are quantitatively trivial.

Cochrane, John. 2012b. Inevitable Slow Recoveries? *The Grumpy Economist*, August 16. **Link**

Cole, Harold L., and Lee E. Ohanian. 2011. Stimulus and the Depression: The Untold Story. *Wall Street Journal*, September 26. **Link**

Council of Economic Advisers (CEA). 2009. Economic Projections and the Budget Outlook. *Council of Economic Advisers Fact Sheets and Reports*, February 28. **Link**

Cowen, Tyler. 2012. When Was It Obvious Our Recovery Would Be So Slow? (Questions That Are Rarely Asked). *Marginal Revolution*, August 17. **Link**

Cushman, David O. 2008. Real Exchange Rates May Have Nonlinear Trends. *International Journal of Finance and Economics* 13(2): 158-173.

Cushman, David O., and Nils Michael. 2011. Nonlinear Trends in Real Exchange Rates: A Panel Unit Root Test Approach. *Journal of International Money and Finance* 30(8): 1619-1637.

Davis, William L., Bob G. Figgins, David Hedengren, and Daniel B. Klein. 2011. Economics Professors' Favorite Economic Thinkers, Journals, and Blogs (along with Party and Policy Views). *Econ Journal Watch* 8(2): 126-146. **Link**

DeLong, J. Bradford. 2009. Permanent and Transitory Components of Real GDP. *Brad DeLong's Semi-Daily Journal*, March 3. **Link**

DeLong, J. Bradford. 2011. Yes, Recovery in the Great Depression Started with Roosevelt: Stephen Williamson Takes to His Fainting Couch Department. *Brad DeLong's Semi-Daily Journal*, September 27. **Link**

Diebold, Francis X. 2008. *Elements of Forecasting*. 4th ed. Mason, Ohio: South-Western.

Elliott, Graham, and Allan Timmerman. 2008. Economic Forecasting. *Journal of Economic Literature* 46(1): 3-56.

Elliott, Graham, Thomas A. Rothenberg, and James H. Stock. 1996. Efficient Tests for an Autoregressive Unit Root. *Econometrica* 64(4): 813-836.

Elmendorf, Douglas W. 2009. Letter to the Honorable Judd Gregg. February 11. **Link**

Enders, Walter. 2004. *Applied Econometric Time Series*. 2nd ed. Hoboken, N.J.: John Wiley and Sons.

Gonçalves, Silvia, and Lutz Kilian. 2004. Bootstrapping Autoregressions with Conditional Heteroskedasticity of Unknown Form. *Journal of Econometrics* 123(1): 89-120.

Greene, William H. 2003. *Econometric Analysis*. 5th ed. Upper Saddle River, N.J.: Prentice Hall.

Hansen, Bruce E. 2007. Least Squares Model Averaging. *Econometrica* 75(4): 1175-1189.

Harris, David, Stephen Leybourne, and Brendan McCabe. 2007. Modified KPSS Tests for Near Integration. *Econometric Theory* 23(2): 355-363.

Hyndman, Rob J., Anne B. Koehler, J. Keith Ord, and Ralph D. Snyder. 2008. *Forecasting with Exponential Smoothing*. Berlin: Springer.

Kim, Chang-Jin, and Charles R. Nelson. 1999. Has the U.S. Economy Become More Stable? A Bayesian Approach Based on a Markov-Switching Model of the Business Cycle. *Review of Economics and Statistics* 81(4): 608-616.

Koop, Gary, and Simon Potter. 2003. Forecasting in Large Macroeconomic Panels using Bayesian Model Averaging. *Federal Reserve Bank of New York Staff Reports* No. 163. Federal Reserve Bank of New York.

Krugman, Paul. 2009. Roots of Evil (Wonkish). *The Conscience of a Liberal, New York Times*, March 3. **Link**

Krugman, Paul. 2011. Bad Faith Economic History. *The Conscience of a Liberal, New York Times*, September 27. **Link**

Larson, Jeff, Christopher Flavelle, and Ryan Knutson. 2010. Stimulus Speed Chart. *ProPublica*, July 30. **Link**

Lütkepohl, Helmut. 2005. *New Introduction to Multiple Time Series Analysis*. Berlin: Springer-Verlag.

Mankiw, N. Gregory. 2008. Stimulus Spending Skeptics. *Greg Mankiw's Blog*, December 18. **Link**

Mankiw, N. Gregory. 2009a. Is Government Spending Too Easy an Answer? *New York Times*, January 10. **Link**

Mankiw, N. Gregory. 2009b. Team Obama on the Unit Root Hypothesis. *Greg Mankiw's Blog*, March 3. **Link**

Mankiw, N. Gregory. 2009c. Wanna Bet Some of That Nobel Money? *Greg Mankiw's Blog*, March 4. **Link**

Mélard, G. 1984. Algorithm AS 197: A Fast Algorithm for the Exact Likelihood of Autoregressive-Moving Average Models. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 33(1): 104-114.

Morley, James C., Charles R. Nelson, and Eric Zivit. 2003. Why Are the Beveridge-Nelson and Unobserved Components Decompositions of GDP So Different? *Review of Economics and Statistics* 85(2): 235-243.

Murray, Christian J., and Charles R. Nelson. 2000. The Uncertain Trend in U.S. GDP. *Journal of Monetary Economics* 46(1): 79-95.

Ng, Serena, and Pierre Perron. 2001. Lag Length Selection and the Construction of Unit Root Tests with Good Size and Power. *Econometrica* 69(6): 1519-1554.

Office of Management and Budget (OMB). 2009. *A New Era of Responsibility: Renewing America's Promise*. February 26. Washington, D.C.: U.S. Government Printing Office. **Link**

Pantula, Sastry G., Graciela Gonzalez-Farias, and Wayne A. Fuller. 1994. A Comparison of Unit-Root Test Criteria. *Journal of Business and Economic Statistics* 12(4): 449-459.

Perron, Pierre. 1989. The Great Crash, the Oil Price Shock, and the Unit Root Hypothesis. *Econometrica* 57(6): 1361-1401.

Perron, Pierre, and Tatsuma Wada. 2009. Let's Take a Break: Trends and Cycles in US Real GDP. *Journal of Monetary Economics* 56(6): 749-765

Pesaran, H. Hashem, and Yongcheol Shin. 1998. Generalized Impulse Response Analysis in Linear Multivariate Models. *Economics Letters* 58(1): 17-29.

Pindyck, Robert S., and Daniel L. Rubinfeld. 1997. *Econometric Models and Economic Forecasts*. 4th ed. New York: McGraw-Hill.

Pissarides, Christopher A. 1992. Loss of Skill During Unemployment and the Persistence of Employment Shocks. *Quarterly Journal of Economics* 107(4): 1371-1391.

Shelley, Gary L., and Frederick H. Wallace. 2011. Further Evidence Regarding Nonlinear Trend Reversion of Real GDP and the CPI. *Economics Letters* 112(1): 56-59.

Stock, James H., and Mark W. Watson. 2002. Has the Business Cycle Changed and Why? *NBER Macroeconomics Annual* 17: 159-230.

Stock, James H., and Mark W. Watson. 2007. *Introduction to Econometrics*. 2nd ed. Boston: Pearson Education.

Taylor, John B. 2012. Debate and Evidence on the Weak Recovery. *Economics One*, May 2. **Link**

Tsay, Ruey S. 2005. *Analysis of Financial Time Series*. 2nd ed. Hoboken, N.J.: John Wiley & Sons.

Williamson, Stephen. 2011. Hal Cole and Lee Ohanian Are Bad Guys. *Stephen Williamson: New Monetarist Economics*, September 27. **Link**

Wright, Jonathan H. 2008. Bayesian Model Averaging and Exchange Rate Forecasts. *Journal of Econometrics* 146(2): 329-341.

About the Author

David O. Cushman is Captain William McKee Professor of Economics and Business at Westminster College (Pennsylvania), and currently visiting professor (nonresident) at Oxford Brookes University. He received a Ph.D. from Vanderbilt University. Cushman previously held regular positions at the University of New Orleans, the University of South Florida, and the University of Saskatchewan, where he is Professor Emeritus of Economics. He has also held visiting positions at the University of North Carolina and Vanderbilt University, and been visiting scholar at Tilburg University and the European University Institute. His most highly cited papers have appeared in the *Journal of Monetary Economics*, *Review of Economics and Statistics*, *Journal of International Economics*, *Review of World Economics*, and *Journal of International Money and Finance*. Cushman also performs on the oboe and earlier in life played professionally with the Daytona Beach Municipal Band, Florida Symphony, and Nashville Symphony. His email is cushmado@westminster.edu.

Go to Archive of Economics in Practice section
Go to September 2012 issue

EJW
ECON JOURNAL WATCH
Scholarly Comments on Academic Economics

ECON JOURNAL WATCH 9(3)
September 2012: 350-365

Rating Government Bonds: Can We Raise Our Grade?

Marc D. Joffe[1]

LINK TO ABSTRACT

In John Moody's inaugural analysis of *Government and Municipal Securities*, published in 1918, the Aaa rating was described as follows: "Bond obligations, whether of countries or municipalities, which are given this rating, are to be regarded as of a high class type of investment and should be considered as entirely secure both as to principal and interest" (10). Since their inception, ratings have been intended to convey risk information to investors. More specifically, ratings are supposed to indicate the probability of default.

In the 94 years since Moody's first assessed government bonds, social science research methods have evolved substantially—far more than have the methods rating agencies use to analyze public sector debt. If a contemporary economist was charged with the task of estimating default probabilities for government bonds, he or she might address the problem by creating a logit or probit regression model fitted against the characteristics of defaulting and non-defaulting governments historically. Unfortunately, major credit rating agencies do not apply this approach to assessing sovereign and municipal bonds, nor do they use a number of other analytical techniques employed in the academic literature.

Academic and commercial models using binary dependent variable methods are well established for corporate credit risk analysis. Other quantitative techniques, including collateral simulations and cashflow waterfall analysis, have been applied to structured assets. FICO and others have used empirical analysis of consumer behavior to assign credit scores to most American adults. Of the various classes of debt, the category that has received the least attention from modelers is the one with most systemic importance: government bonds.

1. Public Sector Credit Solutions, San Francisco, CA 94111.

As a Senior Director at Moody's Analytics until 2011, I learned the benefits of applying quantitative modeling techniques to credit. Moody's Analytics is one of a number of providers that have created successful, model-based alternatives to corporate credit ratings. Unfortunately, commercial vendors have been slow to apply advanced modeling techniques to sovereign and municipal bonds. Barriers include difficulties in acquiring standardized government financial data and a dearth of academic literature upon which models may be built. Software developers may also be concerned about the controversy surrounding government credit assessments, especially given the criticism that Standard & Poor's faced for its 2011 U.S. downgrade.

Given the systemic importance of government solvency, I believe much more needs to be done. Without accurate default probabilities, markets lack guidance needed to set interest rates in accord with underlying risk. Policymakers and the public receive confusing signals about the danger—or lack of danger—arising from their governments' debt burdens.

This paper will examine current rating agency processes, review academic literature on government credit risk, and discuss a research agenda addressing the need to multiply our measures of the risk of government bonds. This research agenda involves collecting historical default data and creating simulation-based government default probability models—work that my group, Public Sector Credit Solutions, has already started.

Rating agency processes

Rating agencies publish documents that explain the factors considered when assigning ratings to a specific class of instruments. Separate documents may be published for sovereign, state, and municipal bond issuers.

I recently reviewed rating methodology documents for U.S. local government issuers published by Moody's (2009), Standard & Poor's (2012a), Fitch (2012b), and Kroll Bond Rating Agency (2012). In aggregate, I found 170 unique factors considered by the four agencies when evaluating municipal credit. Many of these factors, such as "Predictability," are difficult to quantify in a formal analysis, while others would be expected to exhibit a high degree of multicollinearity, such as "Per Capita Income" and "Median Household Income."

Since municipal bond defaults are relatively rare, an approach relying on a much shorter list of variables should be both possible and desirable. The philosophy of science contains a substantial literature advocating simpler—or more parsimonious—models (see Forster and Sober 2004 for a detailed discussion of these issues).

Routine use of a methodology that relies on a large set of independent variables poses implementation challenges. To monitor their ratings effectively, agencies need to be able to collect and analyze updated data as they become available. In the absence of a multiple regression model, changes to independent variables have to be evaluated by analysts to determine whether they should trigger a rating upgrade or downgrade. Available evidence suggests that rating agencies do not effectively execute this monitoring role.

Rating agencies received substantial criticism for their monitoring of residential mortgage-backed securities (RMBS) and collateralized debt obligations (CDO) prior to the financial crisis of 2007 and 2008. The U. S. Senate Permanent Subcommittee on Investigations (2011, 307) found that:

> Resource shortages…impacted the ability of the credit rating agencies to conduct surveillance on outstanding rated RMBS and CDO securities to evaluate their credit risk. The credit rating agencies [CRAs] were contractually obligated to monitor the accuracy of the ratings they issued over the life of the rated transactions. CRA surveillance analysts were supposed to evaluate each rating on an ongoing basis to determine whether the rating should be affirmed, upgraded, or downgraded. To support this analysis, both companies [Moody's and Standard & Poor's] collected substantial annual surveillance fees from the issuers of the financial instruments they rated, and set up surveillance groups to review the ratings. In the case of RMBS and CDO securities, the Subcommittee investigation found evidence that these surveillance groups may have lacked the resources to properly monitor the thousands of rated products.
>
> At Moody's, for example, a 2007 email message disclosed that about 26 surveillance analysts were responsible for tracking over 13,000 rated CDO securities.

These findings relate to structured securities rather than government bonds, so perhaps they are not relevant. On the other hand, it is reasonable to think that if rating agencies under-invested in surveillance for their most profitable asset class—structured finance (Cornaggia, Cornaggia, and Hund 2012)—they have made similar or more egregious under-investments in the surveillance of other asset classes.[2]

2. The *New York Times* reported that, recently, "The sovereign debt team at Moody's [had] about a dozen people" (Creswell and Bowley 2011). The firm rates about 100 sovereigns.

Evidence that ratings do not do a great job of incorporating new information derives from transition matrices published by the firms. Under SEC rules, Nationally Recognized Statistical Rating Organizations (NRSROs) are required to include rating transition matrices as part of their annual regulatory filings. These matrices show the distribution of rating changes over a given period. In my own review of the transition matrices published by Moody's (2012, 25), Standard & Poor's (2012b, table 58), and Fitch (2012a, 10) I found that about 90 percent of municipal bond ratings remain unchanged within a given year.[3] For example, the Standard & Poor's transition matrix (for non-housing municipal issuers) showed that 89.11 percent of AA rated issuers remained AA the following year, while 0.18 percent were upgraded to AAA, 1.62 percent were upgraded to AA+ and a total of 9.09 percent were downgraded to various rating categories ranging from AA− down to BB+.

More interesting are the patterns of the ratings changes. When they occur, upgrades and downgrades are serially correlated, as noted by Nate Silver (2011). If ratings changes were the result of an unbiased process, it would not be possible to predict the future rating trend from the most recent rating change. Silver quotes a Standard & Poor's (2011) report showing that sovereign downgrades were followed by further downgrades in 52 percent of cases and upgrades in only 9 percent of cases over the next two years (there were no changes in the remaining 39 percent of cases during that interval). Upgrades were followed by further upgrades 37 percent of the time and downgrades only 6 percent of the time.

The infrequency and serial correlation of rating changes may be the result of the committee process employed by the firms. Rating changes must be proposed by an analyst, debated, and voted upon at a formal meeting. Rating agency rules clearly state that commercial considerations cannot enter into committee deliberations. The proceedings are confidential, however, and it is impossible for outsiders to determine how participants make their decisions. The serial correlation of rating changes makes us wonder about how political power influences or interacts with the commercial services in question.

Academic research

Limitations on the rating agency processes suggest that an opportunity exists to provide alternatives that predict better. In the matter of corporate bond ratings,

3. Ratings in the transition matrix are underlying ratings that do not reflect the benefits of municipal bond insurance. Prior to the financial crisis, many municipal bonds were rated AAA/Aaa because they were wrapped by policies issued by AAA/Aaa insurers, but credit rating agencies also reported underlying, unenhanced ratings.

a seminal journal article by Robert Merton (1974) eventually launched an industry dedicated to estimating public firm default probabilities based on the market value of their assets as measured by market capitalization. Since the independent variable changes frequently, default probabilities can be updated daily or even in real time. Bankruptcy risk modeling using financial statement data traces its origins to Edward Altman's (1968) Z-Score, and other early work, and has also been commercialized. Both Moody's and Standard & Poor's either acquired firms that commercialized these methodologies or developed them internally—offering evidence that the incumbent rating agencies recognize the validity and power of these quantitative approaches to corporate default probability assessment.

Most academic efforts to estimate government default probabilities have relied on market pricing. A number of researchers have attempted to derive default probabilities from bond yields or credit default swap (CDS) spreads (Remolona, Scatigna, and Wu 2007; Wang, Wu, and Zhang 2008; Longstaff, Pan, Pedersen, and Singleton 2011). In theory, bond yields should be a function of their issuer's credit risk. More specifically, yields should compensate investors for expected loss arising from a potential default. In the literature, expected loss is defined as the product of default probability and loss given default (LGD). LGD is simply the complement of a bond's rate of recovery, and is also called loss severity.

Theoretical bond yields contain a number of components aside from expected loss. Deepak Agrawal, Navneet Arora, and Jeff Bohn (2004) propose an equation for corporate bond yields that includes the risk-free rate of interest, the level of investor aversion to risk, the bond's maturity date, issuer size (as a proxy for liquidity), and the correlation of the bond's default risk with that of other instruments. Yields may also be affected by call provisions that give issuers the option to redeem their bonds prior to maturity.

With respect to U.S. municipal bonds, a further complexity arises as a result of their tax status. As interest on most municipal bonds is exempt from federal, state, and local income taxation, their yields are not directly comparable to those on taxable securities. Some adjustment to the municipal bond yield must be made in order to make it "taxable equivalent." One approach is to convert the tax-free yield to a taxable yield based on the highest prevailing marginal tax rate, on the assumption that municipal investors are predominantly high-income individuals. Given the complexities of the tax code, the heterogeneity of individual investors, and the participation of institutional investors (with different tax considerations), however, the use of the top marginal rate is but a coarse stand-in. John Chalmers (1998) finds that interest rate differentials between long term US Treasuries and federally insured municipals (which are assumed to have no default risk) are not consistent with the tax benefits available to individuals in the top tax bracket.

In corporate credit markets, analysts often derive default probabilities from CDS spreads rather than bond yields. Credit default swaps are insurance contracts that protect the holder against the impact of a default. If the issuer defaults, the CDS seller (or insurer) pays the buyer of the protection the face value of the bond and takes the bond in exchange. Deriving default probabilities from CDS spreads is easier than using bond yields because bonds have more structural complexities, such as call provisions. The applicability of CDS-implied default probabilities to the municipal market is greatly limited by the fact that CDSs trade against a very small number of municipal issuers.

While most large sovereigns are referenced by CDSs, liquidity in these issues is limited. Kamakura Corporation examined sovereign trading volumes reported by the Depository Trust Cleaning Corporation for late 2009 and 2010, finding that the vast majority of sovereign CDS contracts were traded fewer than five times per day (van Deventer 2012).[4] Five transactions per day falls well short of a liquid market—a fact that should be considered in assessing the information content of sovereign CDS spreads.

For spread decomposition to produce perfectly accurate default probability estimates, fixed income markets must perfectly reflect all available information. By implication, such analysis relies upon the strong form of the efficient-markets hypothesis (EMH) advanced by Eugene Fama (1970)—an idea that has often come under attack (e.g., Summers 1986, Crotty 2011). Most tests of EMH have involved equities rather than bonds. In a 2003 survey of EMH literature, Burton Malkiel (2003) identified only one study addressing bond-market efficiency, and that paper found inefficiency in the pricing of corporate bonds (Keim and Stambaugh 1986).

Since the trading volumes of large-cap stocks are much higher than those of municipal bonds, it is not clear that EMH applies at all to municipal bonds. There is a substantial literature documenting the lack of liquidity and transparency in the municipal bond market (see, for example, Chakravarty and Sarkar 1999; Harris and Piwowar 2006; Ang and Green 2011).

Bonds issued by larger sovereigns and sub-sovereigns do enjoy greater liquidity, but it is not clear that they trade based on their underlying credit fundamentals. In 2011, U.S. Treasury yields fell despite the failure of Congress and the Administration to agree on significant fiscal consolidation measures.

Even in the most liquid markets, bubbles and busts occur. When bubbles occur, market prices become detached from traditional measures of intrinsic value. Economists and other researchers have demonstrated skill in assessing intrinsic value and identifying the presence of bubble conditions. Robert Shiller (2000 and 2005) identified both the NASDAQ stock bubble and the real estate bubble using

4. Excludes inter-dealer trades.

such measures. Intrinsic-value analysis of financial assets dates back to the pioneering work of Benjamin Graham and David Dodd (1934) and John Burr Williams (1938), who suggested that stock prices could be benchmarked against measures such as the present value of future dividends and enterprise book value. Shiller (2005) considered building cost indices, price to rent ratios, and other benchmarks in support of his thesis of a property price bubble.

If government default probability and other valuation drivers can be derived independently from price, we can run the yield decomposition apparatus described earlier in reverse to determine intrinsic government bond yields and CDS spreads. Robust measures of fair value bond yields and CDS spreads would enable us to identify instances in which government credit markets are mispricing default risk.

Toward econometric modeling of government default risk

The literature contains a number of efforts to model sovereign bond defaults from fiscal, economic and political variables. For example, Paolo Manasse, Nouriel Roubini and Axel Schimmelpfennig (2003) fit a logit model on data from 37 countries between 1976 and 2001. They found the following variables to be statistically significant:

- Total External Debt / GDP
- Short Term Debt / Reserves
- Interest on Short Term Debt / GDP
- External Debt Service / Reserves
- Current Account Balance / GDP
- Openness
- US Treasury Bill Rate
- Real GDP Growth
- Inflation Volatility
- Inflation > 50 percent
- Presidential Election
- Freedom Index
- Lagged Crisis Indicator

Jens Hilscher and Yves Nosbusch (2010) created a logit model on a data set covering 31 countries from 1994 to 2007, which included 28 defaults. The variables they found to be statistically significant were:

- Volatility of Terms of Trade
- Change in Terms of Trade
- Years Since Last Default
- Debt / GDP
- Reserves / GDP

Because these models considered recent defaults prior to that of Greece, they only addressed emerging-market credit crises. It is only very recently—with the onset of the Eurozone crisis—that defaults among so-called "advanced economy" sovereigns were widely thought to be possible. An indication of the formerly prevailing wisdom is that Basel II rules included a zero risk weight on OECD sovereign debt. The zero-risk weighting meant that banks were not required to hold capital against OECD sovereign bonds, reflecting an assumption that they were default-risk free.

The last time that a significant number of government bond defaults occurred in what are now defined as advanced economies was during the 1930s. The most prominent defaults from this period in the wealthier Anglophone nations included:

Issuer	Year
Alberta (Province), Canada	1936
Arkansas (State), U.S.	1933
Australia *(bonds "voluntarily" swapped for lower rate issues)*	1931
New South Wales (State), Australia	1931
New Zealand *(bonds "voluntarily" swapped for lower rate issues)*	1933
United Kingdom *(made partial payment to U.S. on WWI debt)*	1933
United States *(abrogation of the gold clause)*	1933
Sources: *Moody's Government and Municipal Bond Manual*, *New York Times*, *Australia Year Book*, and *New Zealand Year Book* (details available from the author upon request).	

In all seven cases, interest expense as a percentage of total government revenue exceeded 30 percent. Relatively few large governments that reached ratios in excess of 30 percent avoided default, while all states, provinces, and commonwealths that stayed below this threshold continued to service their debts on time and in full.

The interest expense to revenue ratio normalizes a government's debt service burden against the government's capacity to harvest receipts from its tax base. The ratio has strong intuition. Unlike the debt/GDP ratio, it incorporates interest rates and systemic limits on a government's ability to extract tax revenues.

For an elected official the political cost of defaulting is quite high. It represents a serious embarrassment and it restricts the government's ability to finance future deficits through bond issuance. On the other hand, high levels of debt service crowd out spending required by key constituencies. The dilemma was captured in 1931 by Jack Lang, Premier of New South Wales, when announcing the state's default:

> Parliament in New South Wales was faced with an extremely awkward problem. It was committed to pay to oversea [*sic*] bondholders £700,000. The Government itself had not the money. It was informed, however,

> that this amount would be made available for shipment overseas if the Government needed it. Having in mind the reiterated statement that every £ of credit consumed by the Government meant a £ less for circulation among the primary and secondary industries, the Government was faced with a most difficult problem. If we took the £700,000 which the bank offered us, it meant that £700,000 worth of credit would have to be withdrawn from the primary and secondary industries of New South Wales. Default faced us on either hand. We could default, if we chose, to the farming community by withdrawing £700,000 from it, or we could default to our oversea creditors. Having to choose between our own people and those beyond our shores, we decided that the default should not be to our own citizens. (*Sydney Morning Herald* 1931, 11)

In making the default decision, officials balanced the interests of bondholders and voters/taxpayers. Such a tradeoff is represented by the interest expense to revenue ratio.

That ratio may be one of several metrics that could appear in a logistic model of sovereign, state, and provincial default risk. Further modeling will require the collection of more fiscal, political, and economic data from the 1930s and before.

Many resist the use of older data in modeling government default. Their objections are addressed by Carmen Reinhart and Kenneth Rogoff in a book whose title speaks to this very matter: *This Time Is Different* (2009). The authors state in the preface:

> Above all, our emphasis is on looking at long spans of history to catch sight of "rare" events that are all too often forgotten, although they turn out to be far more common and similar than people seem to think. Indeed, analysts, policy makers, and even academic economists have an unfortunate tendency to view recent experience through the narrow window opened by standard data sets, typically based on a narrow range of experience in terms of countries and time periods. A large fraction of the academic and policy literature on debt and default draws conclusions on data collected since 1980, in no small part because such data are the most readily accessible. (xxvii)

While Reinhart and Rogoff have collected an enormous amount of data in support of their research, the data set does not contain total government revenue, total government expenditure, and interest expenditure. Some of these data are available

from the Center for Financial Stability's Historical Financial Statistics, but only for a limited number of countries for a limited number of years.

Public Sector Credit Solutions

To be effective, default models must be built upon large data sets. My group, Public Sector Credit Solutions, is enhancing the data collected by Reinhart, Rogoff, and the Center for Financial Stability. With the support of a research grant from the National University of Singapore's Risk Management Institute, we are creating a database of sovereign defaults, revenues, expenditures, debt levels, and debt service costs. The database will be freely available to anyone without registration.

Previously we collected data on U.S. municipal bond defaults between 1920 and 1939, which were summarized in Kroll Bond Rating Agency's (2011) inaugural municipal bond default study. We have also collected historical data on U.S. state defaults (Joffe 2012a) and Canadian provincial defaults (Joffe 2012b). New technologies have greatly simplified the task of collecting older government finance data. Many older references have been scanned and are available on line at Google Books, HathiTrust, and other websites. Hardcopy books can be photographed in libraries with high-resolution digital cameras rather than merely photocopied. Advanced optical character recognition (OCR) tools such as Abbyy FineReader facilitate the conversion of scanned or photographed images to usable text. When OCR produces unsatisfactory results, data from scanned or photographed documents can be entered by low-cost outsourcing teams based in developing countries. Collaboration with offshore providers has become easier with the advent of Dropbox and other cloud-based file sharing services.

More data will enable researchers to build better regression models, which can then be applied to current financial data to compute the default probabilities for today's government bond issuers. But any logit or probit model run against current fiscal and economic data will miss some important dynamics. Many advanced-economy governments face increases in pension and healthcare costs as populations age. Also, interest rates may remain low as they have for the last several years, revert to historical means, or go even higher, especially if market participants anticipate high levels of inflation. Academic economists are better equipped than rating agency analysts to estimate future social insurance benefit loads and to create reasonable distributions of future interest rates. Researchers can provide scholarly justification for the way they choose to incorporate factors related to policy, demographics, and macroeconomic events.

In recognition of this reality, Public Sector Credit Solutions has published an open-source tool that enables researchers to create and execute multi-year budget

simulations. This tool, called the Public Sector Credit Framework (PSCF), also allows users to benchmark budget simulation results against a single default threshold or against a logit or probit model. In its basic mode the software counts the number of trials in which a user-specified threshold, e.g., an interest expenditure to total revenue ratio greater than 30 percent, is surpassed, divides that by the total number of trials, and presents the quotient as a default probability. Users can also produce agency-style ratings from the system by providing a default probability to rating map.

PSCF is free and open source. Users are welcome to download it, use it, request enhancements and implement them. PSCF models are also fully transparent so that they can be readily shared with and criticized by other members of the community. The Windows-based software and sample models are available on the PSCF web page (**link**). The source code, along with instructions about how to run a portion of the system under Linux, is available on an open source repository (**link**).

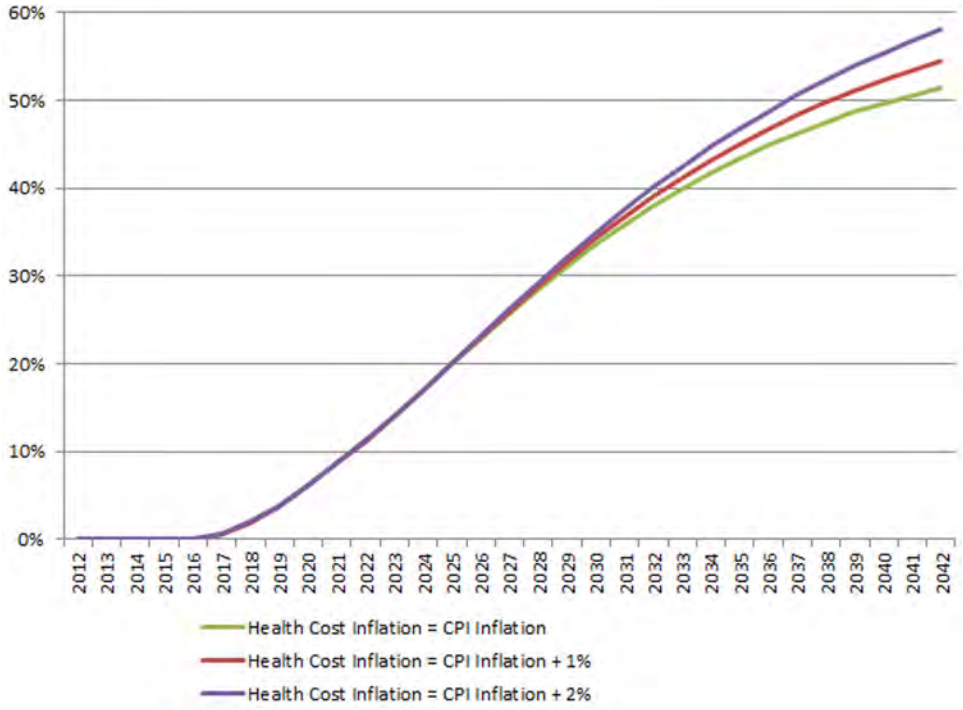
Along with the software, we released sample models for the U.S. Federal Government and for the state of California. Subsequently, we released a model for Italy. All of these samples are available on the PSCF web page.

The sample models provide implementation guidance, but the software itself is quite flexible. You can create any series you want, calculate future revenues and expenditures in any manner you deem appropriate, and use any default threshold ratio you wish. The core services offered by the software are random number generators (supporting uniform, normal and Cauchy/Lorenz, i.e., fat tail, distributions), the generation of simulations parameterized through an Excel front end, and simulation comparisons based on the user-supplied default threshold. A more advanced alternative allows the user to override the single ratio threshold with a C language code block that implements a probit or logit model. While the basic mode generates a binary default/no-default flag for each trial each year, the advanced option returns annual default probabilities between zero and one for each trial that are then averaged across all trials at the end of the simulation run.

With respect to forecast series, the U.S. and California sample models implement a demographic process based on the Lee-Carter mortality model (Lee and Miller 2001). Labor force participation and productivity growth are modeled as autoregressive processes. Random numbers generated in the framework are used to shock forecast rates of annual birth, death, labor force participation, and productivity. GDP growth is then derived from changes in the working age population, labor force participation, and productivity growth. GDP growth forecasts generated by the models tend to be lower than intermediate term historical averages due to lower workforce growth arising from baby-boomer aging.

Revenues are benchmarked against CBO and Legislative Analyst's Office (LAO) forecasts for the U.S. and California respectively. The projected revenues for each simulation trial are a function of the budget office GDP[5] forecast, the difference between the budget office's forecast and the GDP projected by the model for a given trial, and revenue elasticity of the source with respect to GDP.

Figure. U.S. federal fiscal crisis probability, defined as proportion of trials with (interest / revenue) > 30%.



In the U.S. model, expenditures generally conform to CBO projections for the first ten years. After that, Social Security, Medicare, and Medicaid costs are driven by demographic and inflation dynamics. In the sample, healthcare costs are assumed to rise 1% faster than consumer prices generally. But this assumption can be readily replaced with alternative scenarios. The figure above shows U.S. fiscal crisis[6] probabilities for three different scenarios: (1) health cost inflation equals CPI inflation, (2) health cost inflation exceeds CPI inflation by 1% annually, and (3)

5. In the California model, personal income is used in lieu of GDP, but personal income is represented as a fixed percentage of Gross State Product.

6. In deference to those who contend that an outright U.S. treasury default is impossible, we use the term "fiscal crisis probability" in lieu of "default probability". This terminology recognizes that monetization may take the place of outright default.

health cost inflation exceeds CPI inflation by 2% annually. A white paper supplied on the PSCF software page explains how other series were modeled in the samples.

The models we have provided represent only one recommended approach. The transparency of the models and the free availability of the source code underlying the framework are intended to start a conversation among economists and software architects about how best to approach the matter. My hope is that potential users see PSCF not as "Marc Joffe's government default model", but rather as a starting place to create their *own* government default probability models. Further, the open source software itself should become the property of a larger user community that develops and improves it. I welcome an academic department to adopt the open-source project and ask its economics and finance students to improve the software implementation.

Conclusion

As we've seen in Greece and Argentina, sovereign credit crises in relatively high-income nations are disasters. Debt crises force fiscal consolidations. People who are dependent on government salaries and benefits suffer drastic reductions in their living standards. Public sector employees and aid recipients respond with a mixture of despair and protest. When protests become violent, the results are extensive property damage, injuries, and deaths. Since these crises may spread to other OECD sovereigns and sub-sovereigns, the task of estimating their likelihood is an important research question. I invite researchers to explore our historic data and the Public Sector Credit Framework in their pursuit of better tools, better analysis, and better assessments of government credit risk.

References

Agrawal, Deepak, Navneet Arora, and Jeffrey Bohn. 2004. Parsimony in Practice: An EDF-Based Model of Credit Spreads. April 29. Moody's KMV Company (San Francisco). **Link**

Altman, Edward I. 1968. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *Journal of Finance* 23(4): 589-609.

Ang, Andrew, and Richard C. Green. 2011. Lowering Borrowing Costs for States and Municipalities Through CommonMuni. *Hamilton Project Discussion Paper* 2011-01. February. Brookings Institution (Washington, D.C.). **Link**

Chakravarty, Sugato, and Asani Sarkar. 1999. Liquidity in U.S. Fixed Income Markets: A Comparison of the Bid-Ask Spread in Corporate, Government

and Municipal Bond Markets. *Staff Report* 73. March. Federal Reserve Bank of New York (New York). **Link**

Chalmers, John M. R. 1998. Default Risk Cannot Explain Muni Puzzle: Evidence from Municipal Bonds That Are Secured by U.S. Treasury Obligations. *Review of Financial Studies* 11(2): 281-308.

Cornaggia, Jess, Kimberly Rodgers Cornaggia, and John Hund. 2012. Credit Ratings Across Asset Classes: A ≡ A? Working paper. **Link**

Creswell, Julie, and Graham Bowley. 2011. Ratings Firms Misread Signs of Greek Woes. *New York Times*, November 29. **Link**

Crotty, James. 2011. The Realism of Assumptions Does Matter: Why Keynes-Minsky Theory Must Replace Efficient Market Theory as the Guide to Financial Regulation Policy. *Department of Economics Working Paper* 2011-05. University of Massachusetts (Amherst, Mass.). **Link**

Fama, Eugene. 1970. Efficient Capital Markets: A Review of Theory and Empirical Work. *Journal of Finance* 25(2): 383-417.

Fitch Ratings. 2012a. U.S. Public Finance 2011 Transition and Default Study. March 12. Fitch, Inc. (New York). **Link**

Fitch Ratings. 2012b. U.S. Local Government Tax-Supported Rating Criteria. August 14. Fitch, Inc. (New York). **Link**

Forster, Malcolm, and Elliott Sober. 2004. How to Tell When Simpler, More Unified, or Less Ad Hoc Theories Will Provide More Accurate Predictions. *British Journal for the Philosophy of Science* 45(1): 1-34.

Graham, Benjamin, and David Dodd. 1934. *Security Analysis*. New York: McGraw Hill.

Harris, Lawrence E., and Michael S. Piwowar. 2006. Secondary Trading Costs in the Municipal Bond Market. *Journal of Finance* 61(3): 1361-1397.

Hilscher, Jens, and Yves Nosbusch. 2010. Determinants of Sovereign Risk: Macroeconomic Fundamentals and the Pricing of Sovereign Debt. *Review of Finance* 14: 235-262.

Joffe, Marc. 2012a. The Safety of State Bonds: A Historical Perspective. LearnBonds.com (New York). **Link**

Joffe, Marc. 2012b. *Provincial Solvency and Federal Obligations*. Ottawa: Macdonald-Laurier Institute. Forthcoming.

Keim, Donald B., and Robert T. Stambaugh. 1986. Predicting Returns in Stock and Bond Markets. *Journal of Financial Economics* 17: 357-390.

Kroll Bond Rating Agency. 2011. An Analysis of Historical Municipal Bond Defaults. Kroll Bond Rating Agency, Inc. (New York). **Link**

Kroll Bond Rating Agency. 2012. U.S. Local Government General Obligation Rating Methodology. Kroll Bond Rating Agency, Inc. (New York). **Link**

Lee, Ronald, and Timothy Miller. 2001. Evaluating the Performance of the Lee-Carter Method for Forecasting Mortality. *Demography* 38(4): 537-549.

Longstaff, Francis A., Jun Pan, Lasse H. Pedersen, and Kenneth J. Singleton. 2011. How Sovereign Is Sovereign Credit Risk? *American Economic Journal: Macroeconomics* 3(2): 75-103.

Malkiel, Burton G. 2003. The Efficient Market Hypothesis and Its Critics. *Journal of Economic Perspectives* 17(1): 59-82

Manasse, Paolo, Nouriel Roubini, and Axel Schimmelpfennig. 2003. Predicting Sovereign Debt Crises. *IMF Working Paper* 03/221. International Monetary Fund (Washington, D.C.). **Link**

Merton, Robert C. 1974. On the Pricing of Corporate Debt: The Risk Structure of Interest Rates. *Journal of Finance* 29(2): 449-470.

Moody, John. 1918. *Moody's Analyses of Investments. Part III: Government and Municipal Securities*. New York: Moody's Investors Service.

Moody's Investors Service. 2009. Rating Methodology: General Obligation Bonds Issued by U.S. Local Governments. October. Moody's Investors Service, Inc. (New York). **Link**

Moody's Investors Service. 2012. Annual Certification of Form NRSRO 2011. March. Moody's Investors Service, Inc. (New York). **Link**

Reinhart, Carmen M., and Kenneth S. Rogoff. 2009. *This Time Is Different: Eight Centuries of Financial Folly*. Princeton: Princeton University Press.

Remolona, Eli M., Michela Scatigna, and Eliza Wu. 2007. Interpreting Sovereign Spreads. *BIS Quarterly Review*, March: 27-39. **Link**

Shiller, Robert. 2000. *Irrational Exuberance*. 1st ed. Princeton: Princeton University Press.

Shiller, Robert. 2005. *Irrational Exuberance*. 2nd ed. Princeton: Princeton University Press.

Silver, Nate. 2011. Why S&P's Ratings Are Substandard and Porous. *Five Thirty Eight, New York Times*, August 8. **Link**

Standard & Poor's. 2011. Default, Transition, and Recovery: Sovereign Defaults and Rating Transition Data, 2010 Update. February 23. Standard & Poor's Financial Services LLC (New York). **Link**

Standard & Poor's. 2012a. Request for Comment: U.S. Local Governments: Methodology and Assumptions. March 6. Standard & Poor's Financial Services LLC (New York). **Link**

Standard & Poor's. 2012b. Ratings Performance for Exhibit 1 Form NRSRO. July 10. Standard & Poor's Financial Services LLC (New York). **Link**

Summers, Lawrence. 1986. Does the Stock Market Rationally Reflect Fundamental Values? *Journal of Finance* 41(3): 591-601.

Sydney Morning Herald. 1931. Mr. Lang's Defence. April 2, page 11. **Link**

U.S. Senate Permanent Subcommittee on Investigations. 2011. *Wall Street and the Financial Crisis: Anatomy of a Financial Collapse*. Washington, D.C.: Permanent Subcommittee on Investigations.

van Deventer, Donald R. 2012. Sovereign Credit Default Swap Trading Volume. *Kamakura Blog*, January 12. **Link**

Wang, Junbo, Chunchi Wu, and Frank X. Zhang. 2008. Liquidity, Default, Taxes and Yields on Municipal Bonds. *Journal of Banking and Finance* 32(6): 1133-1149.

Williams, John Burr. 1938. *The Theory of Investment Value*. Cambridge, Mass.: Harvard University Press.

About the Author

Marc Joffe is principal consultant at Public Sector Credit Solutions in San Francisco. Until 2011, Joffe was a Senior Director at Moody's Analytics, where he worked for nine years. He researched and coauthored Kroll Bond Rating Agency's 2011 U.S. Municipal Bond Default Study, and he recently published an open-source Public Sector Credit Framework for estimating government bond default probabilities. Prior to joining Moody's, Marc held management and consulting positions at several money center banks in New York and London. He earned his B.A. and MBA from New York University, and he is completing his MPA at San Francisco State University. His email address is marc@publicsectorcredit.org.

Go to Archive of Watchpad section
Go to September 2012 issue