Transcript of 2023 Thomas Laubach Research Conference Session #2
May 19, 2023

TREVOR REEVE. Welcome back to the rest of our program. Before we kick it off, I did want to just take one moment to acknowledge the passing of another of our dear colleagues, Dave Small. Dave Small sadly passed away a couple of weeks ago following the completion of a tremendous 40-year career at the Federal Reserve. He was also one of the founding members of Monetary Affairs with Don Kohn in the back, and we will truly miss him. But I wanted to make sure that we acknowledge that here today. Dave played a very instrumental role, a fundamental role in the completion of the oral history project that I mentioned earlier in a story about Paul Volcker, and the oral history project provides just a wealth of information, a stunning amount of information told firsthand from policy makers and from those who were closely involved in policy decisions. So, thank you for that. Let's now turn to our next session, and I am pleased to introduce David López-Salido, who will chair this session on monetary policy and inflation. And David and I work very closely together. He's an associate director here in the Division of Monetary Affairs and probably needs no introduction to most of you. Thank you.

DAVID LÓPEZ-SALIDO. Thank you, Trevor, and thank you, everybody. This is exciting. We're going to move to the next paper that is on a topic that's very close to Thomas' research as well, because it's going to touch on something that Thomas worked with John and with Rochelle, actually, in thinking about the implication of monetary policy, not just for the aggregate, but also for the different sectors in the economy and the extent to which, in thinking about different sectors, have implications for how to use, to think about monetary policy. So today we have Dave Baqaee that is going to be presenting his work on the supply effects of monetary policy. David is a very [inaudible] and needs almost no introduction. He's one of the stars of his [inaudible]. We are actually privileged to have him for a few weeks in monetary

studies before he moved after graduating from Harvard to LSC, and now he's associate professor at UCLA. And then the discussant will be Rochelle. Everybody knows Rochelle is a Deputy in the division, is an esteemed colleague, and has fantastic academic credentials. So I hope this is going to be good for everybody. So David, whenever you're ready.

DAVID BAQAEE. Great. Well, thanks a lot. It's a real pleasure to be here today. Thanks for inviting me to participate in this conference, which has been a real testament to Thomas' influence and legacy in our profession. So the paper I'm presenting today is called the Supply Side Effects of Monetary Policy. Hopefully these slides will come up soon. Oh, I get to control it. I have the clicker. Thank you. Okay perfect. So this is joint work with another beloved colleague who passed away in 2020 in a very untimely way, Emmanuel Farhi, and a graduate student at Harvard who is exceptional called Kunal Sangani, who I'm sure you're going to hear more about in the future. So whereas the first session that we saw was thinking about how traditionally non-monetary things like demographic change and globalization and long-run structural things like that can have effects on monetary policy, the talk today is going to turn that on its head a little bit, and think about how monetary policy can affect things that are traditionally thought to be non-monetary phenomena.

Specifically in this paper, what we're going to focus on is whether or not there's reasons to think that aggregate demand shocks, which are induced by let's say a monetary policy maker, can have effects on the economy's aggregate productivity. Now the traditional view here, I think it's fair to say, is that the answer is no. Aggregate productivity is determined by structural primitives of the economy like technological progress and R&D and investment activity and those sorts of things, and those are long-run things that are not going to respond to monetary policy very easily. Now, if you adopt this view, there's something awkward in the data that you

then have to explain, which is that our measures of aggregate productivity look like they respond to monetary shocks. And they also look like they respond to things like fiscal stimulus, for example. And the traditional explanation for why this happens is mismeasurement. We just don't know how to measure aggregate productivity properly, so somehow, it's contaminated, and it's a mirage in the data that we see this relationship.

This paper offers an alternative perspective here, and we try to argue that if you -- actually if you write down a sort of standard model but you allow for some realistic firm level of heterogeneity, and I'm going to be very clear about what I mean when I say realistic firm level of heterogeneity, then there's actually a very good reason we should expect monetary policy shocks and other aggregate demand shocks to move aggregate productivity around.

So here, just like a motivating slide to get us started. Hopefully, you can kind of see what I'm showing you here. So here I'm just plotting as motivation an aggregate productivity series. This is multi-factor productivity in red and GDP growth rates per capita for the U.S. over time. And what I want you to focus on is the first thing that you see when you see these two curves is that they tend to move together, and the other thing that you see is that the relationship also seems to have slightly changed over time, in the sense that in the beginning part of the sample sort of from 1950 to about the Volcker era, you've got this very, very strong relationship between TFP and real GDP where there's these -- they almost are moving together. Every time there's a recession, TFP is falling, and then it recovers in the booms. And this is the era during which also real business cycle models were being developed where people looked at the TFP series, and they thought, oh, wow, this could really explain what's going on in terms of cyclical fluctuations. But then post Volcker, this relationship is still there, but it's a lot less strong. And so, hopefully, this paper is able to provide an explanation a little bit for what's going on. What

we're going to propose in this paper is that a lot of these cyclical fluctuations that you see in the early sample are going to be caused by monetary shocks or negative aggregate demand shocks that the monetary authority is not responding to properly. And so to the extent that the Fed is getting better at offsetting negative aggregate demand shocks or not causing negative aggregate demand shocks, then this relationship should be getting weaker over time.

Okay. So now the question is why might a monetary expansion affect aggregate productivity? And the reason that we offer in this paper has to do with allocative efficiency in the economy. So specifically, we're going to write down a model where every time the Central Bank decides to simulate the economy, that's accompanied by a positive supply shock, a positive aggregate supply shock. Why? Because the monetary expansion will induce beneficial reallocations that move the economy closer to the efficient frontier. And that's the thing that's going to generate the boom in productivity coinciding with the boom in output and employment.

Why might this happen? In order for this to happen, two things have to be going on in the economy. The first one is that the initial allocation of resources has to be inefficient in some way, because if the initial allocation of resources is efficient, then a monetary expansion, even if it can shuffle resources across firms, it's not going to have any productivity effects, because the initial allocation was efficient anyway, and there's sort of an envelope theorem that tells you reallocating things is not going to be beneficial. So the first thing we need is that the initial allocation has to be inefficient across firms. The second thing that we need is that different firms have to systematically respond differently to monetary policy shocks or other aggregate demand shocks. Specifically, if you want this channel that I'm going to talk about to operate, you need a monetary expansion to reallocate resources from low marginal revenue product firms to high marginal revenue product firms. Now both of these channels are missing in the traditional kinds

of New Keynesian models that people often work with. Those models are oftentimes linearized at a point where there's no cross-sectional misallocation, so you're never going to have a TFP effect in those models to a first order, and furthermore, those models have this feature that all firms respond in the same way, modular price stickiness. All firms are going to respond in a similar way regardless of their marginal revenue products.

However, the model that we're going to develop today and that I'm going to try to convince you is a good model and has two features, which allow both one and two to be true, and that therefore results in this extra new channel for how monetary policy affects aggregates. The first thing that the model we write down is going to have is firms are going to have variable markups that are different in the cross-section. So firms persistently want to charge different markups. They don't all want to charge the same markup. The second thing that we're going to allow for is that firms have differing degrees to which they pass through cost shocks into their prices, and both of these things are things that have been documented in the literature and industrial organization, and in the literature and international, where people look at these things like pass-throughs. They've seen that firms have different markups, and they have different pass-throughs of marginal costs into their prices. Specifically, the kind of mechanism that's going to operate in our model is the high markup firms in the economy are going to have lower desired pass-throughs, whereas the low markup firms in our economy are going to have high desired pass-throughs. So the way you can think about this is if your firm is charging very low markups to begin with and you get an opportunity to pass through a marginal cost shock into your price, you're going to do that. Whereas if you are a very high markup firm, and you get an opportunity to pass the marginal cost shock into your prices, you might choose not to do that. And that's

going to be enough to generate this kind of expansion pattern that I described earlier, and that's going to result in aggregate productivity moving every time there's a monetary shock.

So in our model, there's going to be this additional transmission mechanism for monetary shocks into the real economy that operates through aggregate productivity, and it's going to generate -- it's going to work by generating basically a positive aggregate supply shock, or if you want, a beneficial cost push shock in the Phillips curve every time there's a monetary expansion. So what are the implications of this? Well, the first implication is that if you have these negative aggregate demand shocks, then there are going to be -- they're going to result in contractions in aggregate TFP that have nothing to do with technological regress. So you can have a recession where aggregate TFP falls, but it doesn't have to do with us forgetting knowledge of how production works. The other thing that the model is going to do is it's going to tie measures like TFPR dispersion to recessions. It's going to be the case in our medical that TFPR dispersion in the cross-section of firms rises every time there's an aggregate demand-driven recession, but it has nothing to do with an increase in uncertainty or anything like that in the model. It's just a flip side of the reduction in aggregate TFP. This mechanism is also going to amplify the impact, and it's going to change the persistence of monetary shock, so it's going to sort of result in a more non-neutral economy, and the effect is strong enough to be comparable quantitatively, we think, to the real rigidities channel for increasing monetary non-neutrality. So it seems like it's a quantitatively significant mechanism.

And finally, what the implication of this analysis is, is that things that we normally think of as being independent to monetary policy, like industrial concentration, are going to have an effect on the transmission mechanism of monetary policy. So if industrial concentration is moving around, then that's going to affect the degree of monetary non-neutrality that your

economy is going to have. Okay. So just to give you a high-level summary of what we're able to do, so we're going to develop a simple model that I hope is going to be relatively transparent and can be put into bigger models by people who are interested in sort of this mechanism. We're going to have a simple four-equation as opposed to three-equation New Keynesian model, and the fourth equation in our model is basically going to be a term that determines how aggregate TFP is endogenously responding to monetary shocks. This model is going to, basically these endogenous TFP movements are going to show up as an endogenous cost push term in the New Keynesian Phillips curve. And they're going to allow the model to generate hump-shaped responses, and they're going to be fairly powerful. So a monetary shock on impact is going to be 44 percent more expansionary on output than in a model that's sort of calibrated with the same parameters but has the supply side mechanism turned off, and it's also going to make monetary policy more persistent than the traditional three-equation New Keynesian model. And then what I also want to do at the end of the talk is provide you with some evidence showing that the predictions of the model about the macro and more importantly the micro evidence concord with the data. Okay. Great. I don't have a ton of time, so with due deference to forebearers, I'm going to just go ahead and into the talk.

So I'm going to start off by giving you a static version of the model that we can solve all the way with pen and paper and hopefully get some intuition, and then once we've seen that, then I'll show you briefly what the dynamic model looks like and what the quantitative results are. And then I'll talk about this empirical evidence that I'm promising. Okay. So how does the static model work? So the static model is going to work like the simplest kind of New Keynesian, two-period New Keynesian model you could write down with one ingredient being different. So we're going to have monopolistic competition in the usual way, but we're going to

move away from CES demand. So CES demand, the feature that it has is when you use CES you're imposing that all firms have the same desired mark-ups, and all firms have the same desired pass-through of marginal cost into their prices. And by relaxing CES, by allowing these more general demand curves, we're going to be able to break those two simplifications. So we're going to generalize the CES assumption using what's known as a Kimble demand aggregator, and then we're going to allow for heterogeneity in our economy, which is going to be very important. So we're going to have firms that have different marginal costs of production, different markups in steady state, and different pass-throughs, and then we're going to layer some Calvo-type frictions on top and just see what happens.

So here's how the model works in period zero. Everybody's sort of in steady state maximizing. There's nothing unexpected happening, and then between period zero and period one, there's this unexpected, if you like, MIT shock where the central bank or some monetary authority changes a nominal price. Now following that nominal shock, the new equilibrium has to be established, and this is where the Calvo comes in. Some firms can adjust their prices and they do, but some firms are stuck. Whatever price they said in zero given their expectations, they now have to live with. That's the source of monetary non-neutrality. Okay so that's it.

So let's try to analyze a model like this. So the households have a typical sort of objective - they're consuming. So, big Y here is GDP consumption output, and L here is employment, since they have some disutility of labor, and then the important thing for us is that the output here is not some CES aggregator over varieties. It's given by this function here. So this *upsilon theta* here is a -- so *theta* is a variety of a good and *upsilon theta* is going to be allowed to vary both with the identity of the variety, and its shape is allowed to be non-CES, right? So if I made *upsilon* just like a power function, this would be CES, but this is going to allow us to move

away from that. And then there's a budget concern in the usual way. So what's important here is that the demand curve for each firm here is now going to take a shape that looks like this, where the quantity that firm *theta* can sell relative to the aggregate depends on the price that firm *theta* can charge relative to some aggregate price index. So competition is mediated through this price index, but crucially for us this *upsilon* function is very general. So you could think of *upsilon* prime as giving you the shape of the residual demand curve, and a traditional analysis forces this to be a power function with a single common elasticity. I'm now allowing it to be whatever function I want, and this is going to be very important for us. Okay. So the firm side is very simple. Firms are just going to -- some fraction of firms are sticky. Some fraction are flex. So with probability 1 minus *delta* you get to set your price before knowing what the nominal shock was. With probability *delta*, you get to set it after knowing the nominal shock. Production is linear in labor. Your marginal cost of production is the nominal wage divided by some productivity shifter that's specific to you as the firm, and you're just going to set your price to maximize profits subject to the residual demand curve that you face.

So how is equilibrium going to work? The bank is going to choose a nominal price. The specific price I'm going to think about because it's going to give me the cleanest equations, but it doesn't make a quantitative difference, is the nominal wage. So the Central Bank here is just going to directly pick the nominal wage. So an expansionary shock here, an inflationary shock here is going to be one in which the nominal wage rises. And then equilibrium has to be established. Everybody is maximizing except the sticky price firms who are stuck. Now think about this monetary shock here, which is moving the nominal wage, and think about how output is going to respond. So output in the model is going to respond for two different reasons. First of all, it could respond because employment rises. This is the traditional kind of Keynesian

demand-side mechanism that we normally think about. But in principle output could also rise because output per worker rises, and so here this is going to be our notion of productivity in this model, and it could be a thing that responds to monetary policy. So this D log A is what I'm going to focus on. Now in order to do that, unfortunately, I have to give you a little bit more notation.

So what's the notation I need to give you? So I'm going to define *sigma* to be the price elasticity of demand faced by variety *theta*. So what's key to point out here is that this price elasticity of demand varies both as a function of how big the firm is. So as you move down the demand curve, the price elasticity is changing, but it also in principle could vary with the identity of the firm itself. So some firms just naturally have a different price elasticity of demand to other firms. The function, it turns out, is given by, you know, the curvature of the *upsilon* function, which is convenient. Now given the price elasticity of demand, every firm has a different desired markup, which is given by its price elasticity of demand in the usual way, but the key here is this is going to now vary as a function of firm size and as a function of the identity of the firm. So that's *mu*. *Mu* is going to play an important role. And then the other thing that's I'm going to have to define is this pass-through object that I talked about, which is impartial equilibrium. If I change the marginal cost of production for the firm, what is the amount by which the firm adjusts its price? So I'm going to call that number *rho*. In a CES model, that number is 1 for every firm, because if your marginal cost goes up by 1 percent, you raise your price by 1 percent. Here's what this thing looks like in the data. So these are estimates from a paper by [inaudible] and co-authors. The little blue lines kind of give you their point estimates, and we just are fitting like a spline through it to show you the pattern more easily. So what they find is that basically for little firms, so this is for the bottom like 80 percent of firms in the sample, the pass-through is like 97

percent. It's almost 100 percent. So it's almost 1 in this notation that I wrote here. So the bottom like 80 percent of firms have a pass-through of basically 1, but as soon as you start bringing big firms into the picture, the pass-through starts to fall very dramatically so that the average pass-through in the data is actually closer to 0.6, rather than 1. So these giant firms have much, much lower pass-throughs than the small firms do. So this is very inconsistent with the traditional CES New Keynesian model, and this is what's going to generate these productivity effects for us. So we have pass-through, markups, and price elasticities. Ah, just a little bit more notation. I'm sorry. I need to take a lot of averages across firms. When I do that, I'm always going to weight by the sales distribution of the firms, and the sales distribution I'm going to call *lambda*. So if you see an expectation with a *lambda*, that means I'm taking an average using *lambda* as the weights. It's a sales-weighted distribution. To give you an idea, if I want the average markup, for example, new bar, I'm going to use the sales-weighted notion rather than just a pure average across firms.

Okay. So now how does aggregate productivity respond in this model? So here is how aggregate productivity is going to respond. It's going to depend on a particular covariance. In particular, it's going to depend on the covariance of the inverse markups and changes in costs or employment in this simple economy. So if it so happens to be that when the monetary shock arrives, resources flow to the high markup firms, then this whole term becomes positive. So the extent to which a monetary shock induces a reallocation of resources from low to high markup firms is the extent to which aggregate productivity rises every time you get a monetary shock. So this is going to be the object of paramount interest, and this thing is something we could go to the data later and directly look at. We could construct this covariance in the cross-section of firms and just look at how it responds to monetary shocks. That's what I'm going to show you later, but this is what you need to be true.

Now why might this happen? Why might this be a positive number? So if we solve out the model -- oh, by the way, I should mention if you take a standard New Keynesian model, *mu theta* is equal to a constant number for every firm, so then this is a covariance with a constant that's always zero. So you don't get anything. So what happens in this model? So here it is solved out in terms of primitives. Every time there's a monetary shock, there is the potential for aggregate productivity to respond, and that potential depends on two covariances. So ignore *kappa rho, kappa delta*. These are just positive constants. Whether this thing is positive or not depends on these covariances. What are the covariances? The covariances are the covariance between desired pass-through and the price elasticity and the covariance between the price stickiness parameter, price flexibility if you like, and the price elasticity. So if either of these covariances in the cross-section of firms is positive, then you're going to get a beneficial increase in aggregate productivity every time there's a monetary expansion. Why is that?

So let's take a look at the first one. The first one is -- sorry, let's take a look at the second one, because it's more mechanical and it's more obvious how it works. Imagine just by luck that the firms that have high price elasticities are also the firms that have very flexible prices. Well, what happens in equilibrium every time there's an inflationary shock? The high price elasticity firms, because they're flexible, are going to adjust their prices. They're going to raise their prices. But the low price elasticity firms, who have lower ability to adjust the prices don't. So you get a change in the relative prices of these firms, which covaries with the initial price elasticity. Remember the price elasticity is related to 1 over the markup. If you're a very price-elastic firm, you have a very low markup, and if you're a very price-inelastic firm, you have a very high markup. So if it's the case that the elastic firms are also flexible, this covariance is positive, and you get an increase in productivity. So this is a kind of a mechanical effect if you

allow the Calvo parameter to covary with price elasticities. This effect is a little more subtle, because this has got nothing to do with the Calvo parameter. This is the covariance between desired pass through and the price elasticity. This operates through the firm's choices itself and the shape of the residual demand curve. This says when the inflationary shock arrives, if the high markup firms are pricing to market and they're very concerned about strategic complementarities, more so than the low markup firms, then they're not going to adjust their prices by as much. That's going to cause the relative price of high markup firms to low markup firms to change, and it will do so in exactly the way that generates a positive covariance here. That gives you the increase in aggregate productivity.

Okay? So I'm going to skip these. So this is the TFP response. Now, what about the output response. If we go all the way down to output, we can see that the response of output can now be decomposed into two terms. One term has to do with the TFP effect, which I mentioned before, and then there's another effect which has to do with the average reduction in markups when you have inflation. This is the traditional Keynesian mechanism, New Keynesian mechanism. It says every time there's an inflationary shock induced by monetary policy, firms don't adjust their prices, so it's as if their markups are being cut. Because it's a reduction in markups, that stimulates labor demand, that stimulates employment. You get a boom caused by the increase in employment. So notice that this mechanism depends on the elasticity, the supply elasticity of labor. So you're in an economy where the labor supply elasticity is zero. This demand side mechanism disappears, of course, because labor is an endowment. So how could you affect it through the Keynesian channel? This mechanism, though, could still operate. Even if the labor supply elasticity was zero in this economy, monetary policy can have an effect on output, and it would do it through productivity rather than through employment. Okay. And of

course the demand side mechanism depends on the traditional sticky price channel and the real rigidities channel, which has to do with strategic complementarities and pricing. Okay. Very good. So this is the static model. Hopefully the intuition is somewhat clear.

So now I'm going to go and develop the dynamic model. How do we do the dynamic model? We do it in the traditional way. We say every firm is setting prices to maximize profit subject to accountable friction. There's a standard consumption Euler equation. Central Bank is following a Taylor rule. Everything is kind of standard except that I'm moving away from CES. That's it. So what happens? The Taylor rule is the usual Taylor rule that everybody knows and loves. The Euler equation, the dynamic IS equation is the usual one. Nothing has changed. But now there's a new term inside the new Keynesian Phillips curve, which I've highlighted in blue. Basically, changes in aggregate productivity are going to show up as a cost push shifter inside the Phillips curve. And the other thing that's going to happen is that the slope is going to be affected by the average pass-through but that's not -- I'm not interested in that because this has nothing to do with heterogeneity. This is the one I'm interested in. Okay. What determines this D log A term, the TFP term? Well, that gets it's own endogenous equation that determines it. So TFP in this model is going to depend on lag TFP and future TFP, but crucially it's going to covary with output, and it's going to covary in output according to this covariance that I mentioned before, which has to do with the pass-through and the price elasticities. So if this covariance is positive every time you get an output boom, TFP also rises, and then it has its own dynamics that have to play out. And it then feeds into the Phillips curve because it's like every time you did your output boom, suddenly there was a beneficial supply shock that came from somewhere that reduces inflation relative to the boom that you got in output and that you might

expect. Okay. So the static model I showed you is a special case of this. If you just impose that the discount factor is zero. So households are kind of completely myopic.

Okay. Very good. So now we want to calibrate this model very quickly. I'm not going to go through the details, but basically we match these pictures. So this is the distribution of firm size. This is the pass-through from other work that's estimated it, and we calibrate the model. I'm going to skip all this and just show you the impulse response functions. So and here we take a model. We feed in, and I see very unfortunately I've forgotten the legend somehow. So here we see it in a monetary shock. So this is a monetary contraction, and we want to see what happens. So there's three lines here. Basically what they are is the -- we've got the green line, the blue line, and the orange Line. The green line is the traditional 3K -- three-equation New Keynesian model. So this is what you get out of Jordi Golly's [phonetic] textbook chapter 3. That's the green line. The orange line is what you get in that model if you allow for imperfect pass-through, but no heterogeneity. So this is the so-called real rigidities mechanism operating through strategic complementaries. The blue line is what happens if you allow for firm level heterogeneity matching the data. And so what do you see? You see that the monetary contraction is way more powerful. It's actually more powerful by almost the same amount that real rigidities boosts the power of monetary policy. So the blue line is way below on the orange and the green, and it -- and this contractionary shock is not as disinflationary. Why? Because in a contraction it's like you get a negative supply shock that raises inflation, and it means you don't get as much disinflation as you would have wanted. What's interesting is the blue model also has predictions for TFP. The blue model says aggregate TFP should fall if there's a monetary contraction, because of the reasons that I explained, whereas the other two models say it should be a flat line, and this contraction in TFP should be accompanied by an increase in TFPR dispersion. Why?

Because if misallocation is getting worse, then TFPR dispersion is getting worse. Markups are becoming more dispersed, because you're moving away from first best. And so both of these are the sorts of things that we want to be thinking about when we go through the data. Okay. So now I've got, I think I'm not, don't have that much time, so I want to speak a little bit about the empirical evidence setting.

DAVID LÓPEZ-SALIDO. Seven.

DAVID BAQAEE. Oh, seven minutes. Okay. No, so I'll show you one more picture then. I've got a little more time. So here I just want to show you this picture because I'm going to try to re-create the same picture again in the data. So here on the left I've just re-plotted the impulse response function for output. So again, the green is the CES model. The orange is a model with real rigidities, but it doesn't have the supply side mechanism operating. The blue is the one that allows all three mechanisms to work together. So the first thing as you saw is on impact having this supply side mechanism boosts the power of monetary policy by a fair amount. The other thing that you can see, though, is it actually also changes the half-life of these shocks. So whereas in the traditional model, the half-life is constant and unaffected by things like real rigidities, in this model because this is a second order -- I'm sorry I got to go all the way back to my -- because this is a second-order difference equation, it actually has the ability to change the half-life of the persistence of these shocks. And so it makes the shocks more persistent as well, and the cumulative effect on output you can see is quite powerful. So, it's again, like I was saying, it's almost, it's actually a little more powerful than real rigidities is, in boosting the responsive monetary policy.

So now, of course, this is kind of the question, the obvious question that comes up is, okay, well, this is very nice but is it consistent with the facts that we see in the data, because this

model is telling a very specific story about cross-sectional patterns you should be able to see across firms. Of course this is very difficult because measuring markups at the level of individual firms is very challenging. We're going to try to provide some evidence to show you that this is plausible. So I'm going to show two kinds of evidence. First, I'm going to show you some aggregate TFP evidence showing that whenever there's contractionary monetary shocks, aggregate TFP is going to respond in the way you expect, but I think, more importantly, I'm going to give you some micro evidence showing that monetary shocks do cause reallocations in the cross-section of firms. Every time you have an exogenous monetary contraction, high markup firms shrink relative to low markup firms. And these reallocation effects don't operate just through first changing their markups differently. They actually operate through employment of resources. Okay. So I'm just going to show you the impulse response functions before stopping.

So here is three different measures of aggregate productivity. So A is labor productivity, output per worker. B is a Solow residual. If you're a big nerd, and you like to think about productivity measurements, there's a thing called the cost-based Solow residual that you should use when there's monopoly power. So for all three of these measures you see that every time there's a monetary contraction, so these are Romer and Romer shocks, but in the paper we do it with high frequency identification and get very similar results. There's a reduction in aggregate TFP, and the reduction in aggregate TFP is about half of what you see in the reduction in output. So almost half of this effect could be explained by the reduction in productivity. Now the more interesting thing for us are the micro-level patterns. So here we use some estimates of firm level markups that come from Gutierrez and Philippon. I can talk about the details of how that measurement exercise is done. So this is in firm level data for public firms in the U.S. So

this is a restricted sample, but nevertheless, and what I'm plotting here is the impulse response for these different covariances. So this is the covariance of markups and changes in costs by firms that are being paid. So you can think of this as a direct measure of resources these firms are using. And this is within three-digit industry codes, and this is just in the aggregate. And so what you see is every time there's a monetary contraction, this covariance is becoming negative, which is to say the high markup firms are losing relative to low markup firms in terms of their utilization of resources, and this is mirrored by a differential response in their markups, which is that the high markup firms are changing their markups differently to low markup firms every time there's a monetary shock, both within industries and in the aggregate picture.

Okay. So I think I'm pretty much out of time, so I'm going to wrap up. So hopefully what's the conclusion I hope you guys come away with? So the first thing is that if you have a distorted economy, then some of the traditional breakdown between supply side at productivity and demand side employment stuff breaks down, and in particular if there's these patterns in the data that I think are realistic, and there's empirical evidence for, then there's this misallocation channel that's operating at business cycle frequencies [inaudible].

DAVID LÓPEZ-SALIDO. Thank you, David.

ROCHELLE EDGE. Okay. So thanks, thanks very much. So before I start, I'd like to thank the organizers for inviting me to discuss David's paper. I really appreciate the opportunity to both to read and discuss this very interesting paper as well as to be part of this conference paying tribute to Thomas. So the conference organizers have done a brilliant job in putting together a conference program that consists of excellent papers and that moreover links to the various research literatures that we all associate Thomas with. Oh, I should -- should I move this along?

DAVID BAQAEE. Yeah.

ROCHELLE EDGE. Oh, yeah.

DAVID BAQAEE. Yeah.

ROCHELLE EDGE. Okay. So the literatures that we all associate with Thomas. So this session's paper, which was a pleasure to read, falls in the area of research on New Keynesian macro models that Thomas both contributed to in his own research and also brought into the way that we undertake monetary policy and macro analysis at the Board. So when we think of the tremendous impact that Thomas had here at the Fed, we typically think of his role as the much loved division director of Monetary Affairs and the much trusted advisor to Fed Chairs Powell and Yellen, but Thomas also spent about a decade of his career here in the Division of Research and Statistics, closely associated with a section called macro and quantitative studies section, which you can think of as the division's macro modeling section.

So during his first spell at the Board between 2000 and 2008, Thomas was an economist in the section, and then after he returned to the Board in 2012, he became a senior officer over the section, and during this time, he made substantial contributions to macro modeling, for which I'm only going to be able to mention a couple. So as an economist, he was one of the first in the section to start working on developing an in-house DSG model that could be used for policy purposes. And then later as the section senior officer, he led an extensive effort to get a large number of prominent DSG models coded up and ready to use for simulations to address practical policy questions. The suite of models continues to be added to and used by macro modelers today to inform policy makers on a range of questions, you know, carrying on this additional legacy of Thomas to policy analysis at the Fed. As the section senior officer, he was also instrumental in making the FRBUS model publicly available, so that researchers outside

of the Board could have access to this model used intensively inside of the Board for policy and economic analysis. So as I said, this session's paper is in the New Keynesian tradition of macro modeling, and to just sort of summarize a bit as well, so what it does is it proposes a new alternative explanation for a well-known and robust stylized fact about total productivity, which is that TFP is procyclical. And this stylized fact has an equally well-known traditional explanation, which is that procyclical TFP reflects mismeasurement.

Now there are various reasons why TFP is mismeasured including perfect competition, increasing returns to scale, but maybe the most straightforward one to explain in words is that inputs into production are mismeasured, that is true inputs are more cyclical than measured inputs for reasons like firms hoarding labor or reducing the work week of capital during downturns, and for these reasons firms hoarding -- sorry, so for these reasons of mismeasured input, this means that TFP ends up being incorrectly measured as procyclical. So this paper takes a different view and instead argues that procyclical TFP is genuine and emerges from two key heterogeneities, a heterogeneity in firm markups and heterogeneity in firm cost pass-through as well as a negative relationship between markups and pass-through. And then what the paper does is develops a New Keynesian model with features that deliver this heterogeneity and thereby procyclical TFP, and it presents empirical evidence to support its alternative explanation. So just a couple of initial reactions. So the alternative explanation for procyclical TFP stemming from firm heterogeneity is an interesting and reasonable one. The model is very clearly explained, and the explanation builds up, you know, very gradually and nicely for the reader from a static to a dynamic version of the model. The paper's appendices consider many variations and extensions to the main model and many empirical robustness checks, and in this respect it's challenging to discuss since any thoughts I had for extensions or robustness checks had already been done.

Indeed, this is probably why I might spend a bit more time on empirical evidence, which is the next bullet. And here I think the paper's empirical evidence is likely the best that can be done given data limitations. So ideally the paper would provide empirical support for its alternative heterogeneous markup pass-through positive explanation that could not be equally interpreted as support for the traditional mismeasurement equation, but in practice this is difficult to do, and I'll talk a bit more about this in my discussion, but I admit that I don't have any good solutions.

So in terms of what I want to cover in my discussion, I want to start off by talking a bit about allocative efficiency. So procyclical allocative efficiency is what generates procyclical TFP and the alternative explanation, and while it was well explained in David's presentation, I want to recap it a little bit. And then I want to say something about the empirical support that the paper provides for the alternative procyclical TFP explanations. Okay. So first allocative efficiency, which emerges from the variety aggregator that David's model ultimately uses. So I thought I would start off by saying something about variety indicators in New Keynesian models, and so this -- so the first green font cell is the familiar CS aggregator, the standard aggregation, the workhorse New Keynesian model, and then all the subsequent green font cells show the various particulars implied by the CES aggregator, and I think most of these were actually in David's first sum notations slides. So you've got the price elasticity of demand, firm's desired markup when it can change its price, firm's desired pass-through, which is in the second row, that is the percent change in the firm's price given the percent change in its marginal cost when it can change its price, which is 1. And then the last green font cell is the quantity of output that firms produce in the steady state and what they produce, that is little y, is less than the efficient level denoted y efficient, and that's, and so basically this is the amount that -- the amount that they underproduce is inversely proportional to the markup. And this

underproduction is, of course, a standard feature in New Keynesian models reflecting their monopolistic competition assumption.

So, as David noted when he was presenting this, there's no heterogeneity with regard to any of the particulars that are described here and with regard to and specifically with -- most importantly, with regard to firms underproducing in this steady state. So even though all the firms are under producing, they're all doing so by the same amount, so there's no misallocation across firms in the steady state. So then what the paper does is it considers a more flexible aggregator, and that's what's shown in orange, and so the first orange cell is the aggregator, the sort of Kimball aggregator, and with this aggregator, the output bundle is defined implicitly by this increasing and concave function of *upsilon*. And then the remaining cells show the particulars for this aggregator where the main thing to note is that the firm's price elasticity of demand and desired markup depend on how much output the firm is producing relative to aggregate output. And then the firm's desired pass-through is more complicated. That's in the second row of expressions, and here notably, it's less than 1, and this incomplete pass-through arises because firms recognize that the price that they set will affect their output relative to aggregate output, and in turn their price elasticity of demand, and so when they take this into account when setting their price, this leads them to less than whatever price change they would otherwise make. Now all firms are identical in the setup. Outside of the steady state their prices and outputs will differ, but that just reflects the differences in the time that they last reset their prices. So this orange last cell is the quantity of output produced in the steady state, and as before it's less than its efficient level, and as before when firms are identical in the steady state, they're all underproducing by the same amount. So again, there's no misallocation across firms with this aggregator. So just to wrap up on the Kimball aggregator, basically the main feature that it adds

is incomplete pass-through. So then the last model feature that the paper adds is to allow for differential markups across in the steady state, so allowing for allocative efficiency in the steady state. And I believe that there's actually, the authors do actually need to go with a more generalized form of the Kimball preferences, though I'm not showing that specifically here. But basically, as can be seen from -- actually I'll go back to this slide, as can be seen from the very last blue font cell in this slide, if steady state markups differ across firms, firms with higher markups will under produce by more than firms with lower markups, and this is basically allocative inefficiency.

Okay. So now if we focus on all the places that markups show up in the various parameters reported in the table, we can see how monetary policy affects allocative efficiency. So just as initially as just noted basically firms with higher markups, and this is in the last cell, firms with higher markups under produce by more. Additionally, and then shifting to the second to last cell, firms with higher markups have lower marginal cost pass-throughs. So when the economy experiences an expansionary monetary policy shock and marginal cost increases, firms with higher markups pass through less of the increase in prices relative to the firms with the lower markups. Now as we know, in sticky price models, firms meet demand at the purser's price, so given this, higher markup firms, since they increase their prices by relatively less, see increase a larger, a relatively larger increase in their output, relatively more than lower markup firms, and because these higher markup firms were the ones who were initially underproducing by more than lower markup firms, an expansionary monetary policy shock reduces allocative inefficiency across firms. So in terms of how the paper gets differential markups across firms in the steady state, it appeals to the fact that there are different sized firms and uses that for their calibration, though they do note that there are other ways to motivate differential markups.

So, procyclical TFP, and this paper's alternative explanation for the phenomena, means procyclical allocative efficiency, and clearly, it's unrelated to actual technology. Indeed, as Baqaee and Farhi note in their 2022 paper in which they sort of go into all this in much more detail, you know, changes in aggregate TFP are equal to aggregate, the technology part of TFP and allocative efficiency. So I did actually just want to briefly note that there are other models in which allocative efficiency shows up, and here I have in mind models which have non-zero trend inflation but do not assume indexation for prices that cannot reset, and as such, they also have dispersion in prices and production in the steady state. So there may be parallels between this paper's model with firm heterogeneity and with Trend inflation. So this is figure five from the paper with a legend actually copied in, and as you see from this slide, the addition of heterogeneous firms to the paper's model reduces the effects of the monetary policy shock on inflation. And since this addition actually also implies a larger effect on output, so that's the sort of larger effect of the blue lines, it's clearly flattening the Phillips curve.

Okay. So then this next slide is from Ascari and Sbordone, and it shows impulse responses from a New Keynesian model with trend inflation but without indexation for prices that cannot reset. So not allowing for indexation also flattens the Phillips curve, although here higher trend inflation actually flattens the Phillips curve which goes against the sort of trend over history. The main reason I thought to note the paper by Ascari and Sbordone is that the authors conclude their paper by noting that implications for monetary policy is something that they could consider in future work, and Ascari and Sbordone considered the implications of trend inflation without indexation for issues like determinacy, optimal stabilization policy as well as some other issues that are sort of more relevant for trend inflation itself, but their paper could nonetheless be helpful to look at in thinking about monetary policy implications of this paper's model.

Okay. So I wanted to say a couple of things about empirical support starting with the static model's results concerning the flattening of the Phillips curve. So here I mainly wanted to note that a couple of the model's predictions concerning the flattening of the Phillips curve were encountered as some empirical facts. So the charts in this slide show impulse response functions to unemployment rate gap shocks with responses for wage inflation in the top chart and responses to price inflation in the bottom chart, and these impulse responses are generated from a time-varying parameter via our model, which means that it's possible to generate impulse response functions at different points in time. And I'm showing here 1975, which is in black, and 1985, which is in blue, and 1995, which is in orange, and 19 -- sorry 2019 in red. So one point evident from these charts is that while the price Phillips curve has flattened, which is the lower chart, the wage Phillips curve has not flattened. Note also that from the price Phillips curve chart that essentially all of the flattening occurred by 1995. So okay, so this chart shows some charts from the paper. These actually -- these weren't charts that David went through in his presentation. So the top two plot the slopes of the wage and price Philips curve for different specifications in the model and for differing degrees of industry concentration. So as you see, both the wage and price Philips curve flattened going from the CES aggregator, which is the green line, to a model that has real rigidities but doesn't have the misallocation channel, which is shown by the orange line, and then to a model that adds heterogeneous firms and the misallocation line, which is -- sorry misallocation channel, which is shown by the blue line. So these charts also show that increasing concentration that's moving along with the charts, the increase in concentration flattens the Phillips curve. So you can see that from the declining orange and blue lines as the Gini coefficient increases where this represents the employment distribution becoming more concentrated. And the paper puts increasing concentration in some

context by showing -- and these are some appendix charts, the amounts by which between 1978 to 2018 concentration increased for firm employment overall as well as firm employment in the retail sector. So note that the model predicts a flattening in both the price and wage Phillips curve, whereas the impulse response functions that I showed earlier, only the price Phillips curve has flattened. Now David and I actually chatted separately, and he did point out that there's -- in his model they do have less of a flattening in the wage Phillips curve, and he might have some things to add on that later. I guess I'd also note that the firm employment concentration charts show as much, if not more, of the increase in concentration having occurred post 1995, and so I've drawn a little line there for 1995. As occurred before it, whereas for the price impulse response functions that I was showing earlier, all of the Phillip's curves flattening occurred by 1995 and very little afterwards. So, okay, two minutes. Oh, no, yeah. Okay. I have a –

DAVID LÓPEZ-SALIDO. You have four.

ROCHELLE EDGE. Okay. I have a couple of small comments on the empirical evidence section, though as I mentioned for the issues I'm raising, I don't really have good solutions. So as I mentioned at the start, ideally the paper's empirical evidence would provide support for the paper's alternative heterogeneous markup, pass-through, explanations for procyclical TFP that could not be equally interpreted as support for the traditional mismeasurement explanation. The paper's macro level and cross-sector evidence, however, have difficulty doing this. So purified solar residuals that correct for input mismeasurement, perfect competition, increasing returns to scale do exist per research by Busufinald [phonetic] and Kimball, now given this one might think to look at the effects of monetary policy shocks on these purified estimates of TFP, to consider the paper's alternative explanation. However, as the authors note, methods for estimating purified solar residuals assume that residuals are orthogonal

to monetary policy of shocks, and indeed all demand shocks. So if the authors were to use these purified residuals for their macro level empirical evidence, they would find by construction monetary policy having no effect. So for the macro empirical evidence they have to use sort of unpurified TFP estimates. So since these estimates very likely have some procyclicality due to mismeasurement, it's likely not possible to tell whether it is the alternative heterogeneous markup pass-through explanation or the traditional mismeasurement explanation that accounts for the effects of monetary policy on TFP.

DAVID LÓPEZ-SALIDO. Two minutes.

ROCHELLE EDGE. Two minutes? Okay. I think the cross-sectional evidence also uses unpurified TFP, so here too it might also be hard to tell between the two explanations. The micro level evidence on reallocation and markup seems to me to be more directly consider the paper's alternative procyclical explanation. So I actually do not have a concluding slide, but I mainly just wanted to wrap up by saying that this is a very interesting, very clearly explained, and very thorough paper. I liked it a lot. I really enjoyed studying it. So thank you for including me on the program. It was a pleasure to read and discuss David's paper, and I am grateful to be part of this conference remembering and paying tribute to Thomas.

DAVID LÓPEZ-SALIDO. Thank you, Rochelle.

[Applause]

DAVID LÓPEZ-SALIDO. So David, we'll give you five minutes, is that enough?

DAVID BAQAEE. That’s plenty, I think.

DAVID LÓPEZ-SALIDO. Plenty?

DAVID BAQAEE. Yeah, sure. Okay. Well, thank you so much, Rochelle, for those comments. So I agree basically with everything that Rochelle said. Maybe I can just like expand a little bit on some of the issues here. So one of the things -- so Rochelle went right for the thing that I think I would in my view is probably the weakest part of the paper and it's also the part that I didn't really talk about in my presentation, because I think I haven't made up my mind about what's going on basically. So I don't view that as like the most important part of the paper, but it's something that I think is worth talking about, which is the following observation that we made in our model.

So I had slides on it, but I ran out of time I think, so I didn't get a chance to discuss it at all. It's this thing about the concentration and how it affects monetary non-neutrality in this model. So this is the figure that Rochelle is showing. So in this model it's true that changing concentration, industrial concentration, is going to have effect on what I'm tentatively going to call the slope of the Phillips curve, but it's not really the slope of the Phillips curve. So we need to be careful there. What I really should have done here is instead of calling this the slope of the Philips curve, if I was being careful, was to call this monetary non-neutrality in the model, which is not the same thing as the slope of the Phillips curve. So like this is the static model. It's easier to see it in the dynamic model. Where is the dynamic model? I'm sorry for flipping back and forth.

So here's the dynamic model, okay, and in this model the slope of the Phillips curve is going to be affected by industrial concentration to the extent that average pass-through is affected by industrial concentration. So you can imagine that if I make industries more concentrated so the big firms become bigger players in their markets where they're operating, then average pass-through may decline, and that will flatten the slope of the Phillips curve. But

the mechanism that's -- but that's not really kind of the mechanism that I've been talking about, because what I've been talking about is these cost push shocks that show up in the Phillips curve, that operate through productivity, and these ones, their strength is also affected by industrial concentration, but they don't necessarily show up in the slope of the new Keynesian Phillips curve if you were to estimate this object. Now when I solve the static model, I can kind of solve the whole fixed point and write the whole thing as like a single, I didn't show this either, as like a single derivative, which is the change in output and the change in wages, and here it looks like a flattening of the Phillips curve. So it's got -- what I'm trying to say is I think you have to be very careful when you go to the evidence that the way that this stuff is estimated in the data is consistent with the way it would work inside the model. It may or may not be.

Now that's one issue I think that's kind of subtle but is important. There's a different issue, which is more substantive, to do with whether or not this is going on, which is that in my opinion we don't really know whether or not concentration has gone up or down. So it's true that -- so that's why in the paper when we do this exercise where we show how monetary non-neutrality responds to concentration, we do it as a thought experiment rather than as evidence of how the U.S. economy has changed. We put these numbers here just to give you like order of magnitudes, not to say this is what happened in the time series. The reason is because in the concentration literature aggregate trends and disaggregated trends are doing different things. So while at the national level concentration has been going up, which is those pictures in our appendix that Rochelle was showing, at the disaggregated level sometimes concentration is going down so the way you can think about this is like if Walmart is opening more branches in sort of places where it didn't have branches before, at the national level there's more concentration because there's more Walmarts taking a bigger share, but at the disaggregated

level, in every single market where Walmart is entering where it wasn't before, concentration has gone down. And so these kind of disparate effects don't paint a super clear picture in the time series of what's going on. So that's why we were very tentative when we talked about this stuff. I wish we could say something more definitive there.

So that's one issue. The other issue I wanted to talk about, which I also think is super important, is this issue of the purified Solow residual. So it's become standard practice, I think, among people who work with aggregate time series of productivity, to use these purified Solow residuals. Sometimes they're called capacity-adjusted Solow capacity, utilization adjusted Solow residuals. So John Farnell, for example, is somebody who's done a lot of work on this. What happens in our model is these capacity-adjusted Solow residuals they're identifying assumptions for constructing them is actually violated inside our model because usually the way that people estimate these things is they make an assumption that demand shocks, by which I mean monetary and fiscal stimulus, do not have effects on measures of sector level productivity, and so that identifying assumption is what allows those measures to be constructed. In our model those are invalid, so I think it's a super interesting and important question basically to revisit that literature and to try to construct measures of capacity utilization adjusted TFP that are robust to this particular issue, because I think my personal opinion is that both things matter probably. I think -- well I wrote this paper -- so I think the supply side mechanism is there, but I'm not so bold as to say it's everything and there's none of that stuff going on in the data to do with improper measurement of input. I think both are probably going on. Even in our quantitative results you could probably see that, which is that in the data, TFPs respond even more strongly than our model predicts. So that opens the door for potentially for capacity utilization to still play a role, but I think it's a paper that needs to be written, in my opinion, so thanks a lot.

DAVID LÓPEZ-SALIDO. Thank you, David. I think we have time to collect some questions if there are any. Mateo…

MATEO. So there is a couple of other papers that more or less have a similar flavor in the sense that they also have this, you can have Calvo model with the kind of heterogeneous price CNS or a [inaudible] higher order, and it also generates this fact that firms with lower markups have higher pass-throughs. So I was curious and I know you know it because it's cited, so I was curious like at the micro level are there differences in what you should find if your mechanism relies say on Kimball vis-a-vis heterogeneous price stickiness or some other story because you know most of the aggregate responses to test these models will look like that. And I was more curious to know at the micro level whether you can test Kimball say vis-a-vis heterogeneous price stickiness that gives the same results.

DAVID BAQAEE. That's a great question. Thanks. So this is kind of, I think, a helpful slide to look at if you want to think about these different mechanisms. So this slide, which by the way this equation the version of it will also hold for the dynamic model, and the static model is the cleanest way to write it, is you can see that the responsive TFP to these monetary shocks depends on two covariances. One is the covariance of the markup. Here *sigma*, remember, is the price elasticity, so it determines the markup. It's effectively the inverse markup, the covariance of the markup with desired pass-through, and this is what I emphasized in my talk today, and then there's a second mechanism that could give it to you, which is a covariance of the markup with the price flexibility of the firms. Now either way, these two things are going to result in realized pass-through covariant with a markup in a systematic way, where realized pass-through is a combination of desired pass-through and price flexibility. So in that sense these two mechanisms are very similar, and so we are able to kind of like talk about both, but when I think about micro

foundations, I focused on evidence about this one. And Meyer and Reynold, they focus on this one. Now what's the -- now I don't have anything against that story per se. It's actually -- so this -- I can tell you the history of how these papers came to me, but this is something that we were open to as well. Meyer and Reynold kind of gave evidence for it and built the micro foundation, but when we started writing this paper, was before the Meyer and Reynold paper. We just said there's these mechanisms and that we're going to focus on this one rather than this one. Now what about the micro evidence? So the micro evidence that you would marshal to support one story versus the other is different, which is that in the one case you would want to look at things like whether or not markup systematically covary with price flexibility or not, which is something that's necessary for the other mechanism, and the other thing you'd want to look at is when we construct these notions of reallocation measures, you want to think about do these markups that we're measuring do they mostly reflect -- well, no actually, it's really, I have to say, it's the way I think you would do it is you would have to look at the covariance in the cross-section of the markups with the price flexibility parameters, versus for us the cross-sectional covariance between markups and desired pass-throughs. I think that's one way to separate these two stories from one another. So I think there's a -- I would -- my understanding of the literature is there's a lot of robust evidence for the covariance of desired pass-throughs with the level of markups from like the IO and trade side. The covariance between price flexibility and the level of the markup, it's a little more controversial because there's some conflicting evidence, there's some papers where they say big firms are more flexible than small firms, and big firms are the ones that have higher markups. And some papers that say no. But yeah, I think thing in principle the two mechanisms are related to one another, though, yeah.

DAVID LÓPEZ-SALIDO. Any other questions, so Benoît?

BENOÎT. Yeah, so thanks a lot for the fascinating presentation. Can you walk us through the main argument of the paper but precisely taking like a large firm, say Apple, and then the mom and pop local store.

DAVID BAQAEE. Okay.

BENOÎT. And what exactly you mean in terms of --

DAVID BAQAEE. Yeah.

BENOÎT. -- so you know the employees of mom and pop, they have a suddenly a job at --

DAVID BAQAEE. Yeah

BENOÎT. -- Apple, and so you know, how it works in this --

DAVID BAQAEE. Yep

BENOÎT. -- simplified version of the model where you have like -- and so [inaudible] as well as higher markup and is also more efficient. So in this context --

DAVID BAQAEE. Yeah

BENOÎT. -- can you try to explain the --

DAVID BAQAEE. Good question. Because I've been working a lot on issues to do with allocative efficiency, sometimes I take things for granted that I think are not so clear to people who haven't been obsessing about these issues so much. So and there's just like a couple of things I want to clarify. So the first thing is you said let's do an example with Apple and a mom and pop store, and then Apple charges a high markup, and Apple is more efficient, maybe, and so maybe this story has something to do with that. The first thing I want to say is you have to be very careful when you use the word efficient here. The sort of stuff I'm talking about is not about who is more efficient in a quantity sense, because somebody being more efficient than

somebody else in the sense that like let's say my TFPQ, this is to say my quantity of output per unit input that you give me may be higher than somebody else's, that doesn't mean I'm too small or that me getting bigger would be a good thing for efficiency, because if there's no initial distortions, the cross-sectional allocation of resources is efficient, that means that reallocating things from one guy to another will not affect efficiency to a first order because of the envelope theorem, and it will reduce it to a second order, because you're screwing up the allocation. So this has got nothing to do with physical efficiency of production. This is only about the initial distortion that makes some guys too big relative to other guys in terms of how much resources they use.

So in this world Apple is too small if and only if Apple is charging a higher markup than the mom and pop store. If you hold their physical technologies constant and just imagine a world where the mom and pop store is charging a higher markup than Apple, then the mom and pop store is too small, and Apple is the one that's too big. So it's really about a comparison of markups, not a comparison of physical efficiency in production. So that's the first thing. Now in this paper, what's happening is every time there's a monetary expansion, it's like marginal costs are going up. So you can think of it as like let's say the wage is going up, because inflation has happened. Now when the wage goes up, different firms think to themselves, by how much do I want to adjust my price? In this model what's happening is if you're a low markup guy, you adjust your price one for one with your wage, with the marginal cost. Why? Because you're not playing games, if you like, with your price. Your wage -- the costs go up, raise your prices. You have razor-thin margins. Your markup is very low, and so you just raise your price to reflect the increase in the cost. So that would be if you like the mom and pop store here. They have razor-thin margins and when the cost of the stuff they bring in the store goes up, they've raised the

price by the same amount, but if you're a sophisticated guy who has high markups, so you're already charging fat margins, when the costs of production go up, you are more likely to start playing games and thinking to yourself, okay, well what will the market bear? Can I just pass this into my prices one for one or is that going to be very bad for me because it's going to affect the amount of demand that I get and in this model what's happening is the guys who charge the high markups are the ones who don't pass the marginal cost increases into their price as much. That means that they cut their margin somewhat because they're pricing to market. There's some sticky guys who haven't adjusted their prices, and you have to maintain competition with them. So you don't adjust your price because you're saying well these guys aren't increasing their prices because they have sticky prices, so I'm not going to increase my price by as much. So what happens, though, is among the flex guys, the markup of the high markup guy is going to fall, because they're pricing to market, but the markup of the low markup guy is not changing because they passed the cost into their price. But then that means that the high markup guy cut its markup relative to the low markup guy. But that's exactly what you want from an efficiency perspective, because the high markup guy is using too few resources relative to the low markup guy. So if they cut their markups, they get more resources, and that boosts efficiency, the way we measure it.

And so that's why I was showing these pictures, right, to show you exactly how resources flow in the cross-section of firms and how that covaries with the initial markups those firms are charging. Now, in my story and in my calibration, like when I -- I'm out of time. I just want to say one thing. In my calibration, big firm is the firm that charges the high markup, but that doesn't need to be the case. Like sometimes, you know, if you go like you buy fancy wine, the fanciest wine doesn't have the biggest market share. The fanciest wine does have a high

markup though, and it tends to be the one that prices the market. So all of this could be operating in the dimension of firm size, but it doesn't have to be. So hopefully that sort of clarifies some things and all. Okay.

DAVID LÓPEZ-SALIDO. Thank you for these fascinating papers. Thank you, David. Thank you, Rochelle, for the discussion, and we're going to have a break for 15 minutes, and we come back at 2:30. Thank you.

[Applause]