

Not-for-Publication Appendix: Response to Previous Referee Reports for “Through the Looking Glass: A WARPed View of Real Exchange Rate History”

Note to referees: Thank you very much for agreeing to referee my paper. Previous referees have made many insightful comments, and have improved my paper tremendously. To make your job easier, I have attached my previous referee reports along with responses. The first goal of doing so is that, if you previously refereed my paper, you can see my response to your comments without having to dig through my entire paper to see how I addressed your concerns and suggestions. Even if you haven't previously refereed my paper, it may be helpful to see what other referees have written. (For one, if you want to kill the paper without reading it, you could simply regurgitate what others have written.) But my hope is that if you think you have found some obvious, fundamental mistake in the paper, you might find that this concern has in fact been addressed (perhaps somewhere in a footnote or appendix, where you are otherwise unlikely to see it).

For example, one frequent comment I have received so far from referees is that the measures presented in the paper are obviously inferior to the Terms of Trade as measures of competitiveness. If true, this would be a fundamental oversight of this paper (or, at least the original draft – the most recent draft also includes new Terms of Trade indices using PWT data). However, as I show in the current draft of the paper, empirically the Terms of Trade appear to be uncorrelated with trade flows. And, even theoretically, it isn't clear that the Terms of Trade, or measures designed to exclude non-traded goods from consideration, are preferable. One referee even wrote that s/he did not think that *wages* mattered for competitiveness. The reason non-traded inputs, such as wages, belong in measures of competitiveness is because, if in some sectors the world output price is fixed (which could be true of goods such as wheat or for iPhones), then the prices of local non-traded inputs such as labor costs will determine location decisions for firms in these sectors. And in the US case, it appears that the dollar-price stickiness of many goods meant that, in practice, the ToT movements were driven by oil price shocks.

The second most common complaint is that I should provide an explicit mathematical justification for something in the paper, such as the Balassa-Samuelson adjustment, or for the proposition that exchange rates matter for trade. Because these theories are well-known, and because I don't seek to add anything to them, for the most part I have left the theory in the footnotes and appendices, where perhaps they are not as easy for referees to find as they would be if they were in the body of the text.

The third most common complaint has been about the trade-weights, I suspect because this is an intuitively easy area to grasp compared with the discussion of the index numbers and data. While I do make fairly obvious and intuitive improvements to both the IMF and the Fed's trade-weighting scheme (I use the BIS's scheme, also used by Chinn 2006, and I increase the number of trading partners included in the index from roughly 30 to 154), trade-weights are not the focus of this paper and, overall, I concur with the conclusions of previous researchers in this field that further improvements in trade-weights are unlikely to be a promising avenue for further substantial improvement of RER indices. This doesn't mean that the improvements I have made to the IMF's scheme did not make a difference -- the IMF's used fixed trade-weights and excluded China, and so updating to the Fed or the BIS schemes of course make a difference. But improving from the Fed to the BIS scheme, or using trade weights computed from trade in value-

added instead of total trade (a method which cannot be done for all trade weights in any case due to data limitations), somewhat surprisingly makes very little difference to the weights themselves and thus to the indices. This is why many researchers, including myself, have concluded that further improvements would be unlikely to bear much fruit, but they would require rather involved data efforts, possibly including imputing data and making educated guesses. That said, if given an R&R on condition that I make a large time investment to improve the weights even further, I would do it as I do have some ideas about how this could be done. I am also very open to any concrete suggestions referees may have to improve them further, (aside from the referee who proposed I use trade *volumes* rather than values, as prices are a convenient way to compare apples and oranges).

Journal Referee Reports

Dear Douglas,

I have now received the referee reports on your submission. Unfortunately the news is not good. Based on the reports and my reading of the paper, I will reject.

Both referees are expert in this area and both recommend rejection. Their comments on the paper are careful, substantial, and well argued, and in this circumstance I feel I have to concur with their judgment.

Referee 1 is sympathetic to the paper, but doubts that it convincingly achieves its goals. They are not persuaded that the proposed new real exchange rate measures solve the problems any better than the old measures conceptually or empirically, and do not see how the new series do any better in practice in predicting trade flows.

Response: Conceptually, the theoretical advantages of my WARULC series over the IMF's series – which uses fixed trade weights, does not include China, and suffers from 'trading partner substitution bias' – seem relatively straightforward. Empirically, the IMF's series implies that US manufacturing became substantially more competitive through the 1990s – just as it was collapsing, while my series can predict the collapse quite accurately. The IMF's index should predict large trade surpluses in the mid-1990s and late 2000s, whereas the WARULC series predicts large deficits. These differences are not subtle.

However, I do acknowledge that, in the US case, the P-WARP index turned out to be very similar to the WARP index (in terms of changes over time at least – the levels are different), and that the key addition in this portion of the paper was extending WARP to the 1820-2011 period instead of 1970-2006, to show that relative prices had not been as high as they were in 2001 since the worst year of the Great Depression, and to provide these indices publicly for the first time. Theoretically, I think it's easy to make the case that P-WARP is an improvement. Trading more with countries with low price levels will obviously not be a large shock to trade if low price/wage countries also have low productivity – consider an increase in trade with Bangladesh, for example. Additionally, trade economists traditionally have a clear preference for RER indices based on unit labor costs.

Referee 2 places a major critical focus on one key modeling issue: that, as in Fahle et al., you do not provide a clear definition of what competitiveness is, so you cannot set up an aggregation problem that the proposed WARP index solves. They also note that your extension to WARP should be judged relative to a basic WARP benchmark not the IMF REER benchmark.

Response: In the updated draft, I provide a model which shows that relative prices matter and thus that WAR indices should be preferred to Divisia. I now also provide a definition of competitiveness (using a standard dictionary definition). I focused on testing my WARULC vs. the IMF REER index because (1) economists have generally preferred measures of competitiveness based on unit labor costs, and (2) because, as I mention, my P-WARP index is so similar to WARP that I did not think comparing any minor differences between them was interesting, and (3) The WARULC index is just for the manufacturing sector, while WARP and P-WARP are for the economy as a whole. In any case, now I have added these comparisons.

I hope these comments will be useful as you revise the paper further.

JIE referee reports may be submitted either as text or electronic files (such as PDF.) Comments submitted as text appear below my signature. Reports submitted as electronic files can be accessed at the JIE/Elsevier website (ees.elsevier.com/JIE) to which you submitted your paper. A link appears below.

Thank you for giving us the opportunity to consider your work.

Yours sincerely,
XXXXXXXXXX

Reviewers' comments:

Report on ‘Through the Looking Glass’

The goal of this paper is, as stated, to produce ‘theoretically-appropriate measures of real exchange rate indices that an applied researcher could use to gauge the impact of RER movements on trade’. This is a worthy goal since the existing real exchange rate indices suffer from the (potentially) important shortcoming that shifts in trade from high-cost advanced economies to low-cost emerging economies are not reflected in the real exchange rate. At least in part for this reason, officially published real exchange rate indices are given in index format, signalling that only changes in the real exchange rate can be measured using these indices. However, even the rate of change could be biased as shift towards lower-cost trading partners would lead to a decline in the weighted average price that (e.g.) importers would have to pay for a given product. From a pure index-number point of view, this line of argument certainly makes sense and the cited work of Houseman et al. (2011) illustrates that this source of bias in the US input price index leads to biased measures of productivity growth. In the current context, this

bias could be important since, ideally, a real exchange rate should be usable to estimate and forecast changes in trade patterns. That said, I have a number of concerns about the paper. Before I get into those, let me emphasise that the author should be commended for the effort that went into the data construction. Likewise, making these publicly available is useful for the broader profession.

However, my main concern is that I am not convinced that the proposed alternatives, the so-called BS-WARP and WARULC series, are a good solution to the initial problem. The difficult part in solving the problem the author sets out to do is to compare ‘like with like’. If Chinese exporters increase their market share in, say, the US, this suggests that the prices of the products it exports were, in quality-adjusted terms, lower than prices of competitors. But deciding by how much is not straightforward. What we can be sure of is that the average price level of China is not an appropriate measure: prices in China are lower than in the US mostly because non-traded products are cheaper. This disqualifies the WARP series of Fahle et al. (2008).

Response: Two points here. First, this critique also applies to existing RER measures produced by the IMF and the Federal Reserve, the latter of which uses country-specific CPIs as the base. Thus the indices I create could still be improvements even if they are still imperfect and do not solve all the problems with existing indices. Secondly, there isn’t a strong theoretical rationale for excluding all non-traded goods from comparison. Imagine that Apple will charge \$800 for an iPhone in the US because it determines that that is the price that the domestic market will bear, or that a firm operates in a sector, such as grain production, with a fixed world price. Where will the firm produce? It will produce where it is the cheapest to do so. And it will be the cheapest to produce where non-traded inputs, such as labor costs, or the cost of business services, building materials and land or renting office space are the cheapest. In that case, it isn’t at all clear why you would necessarily want to exclude non-traded goods. For sectors such as these, tradable goods prices alone could be very misleading.

But even the BS-WARP that is proposed in this paper is not a clear improvement as it merely adjusts the average price level for the average price level observed in countries with similar income levels. Yes, this is used by Rodrik (2008) as an indicator of exchange rate over- or undervaluation, but I would argue that it is not a precise measure of the over- or undervaluation of every traded product.

Response: First, I agree that Balassa-Samuelson residuals aren’t a perfect measure (and I certainly did not mean to imply that it is a perfect or precise measure, but I would argue that if we limit ourselves to RER indices which are a “precise measure of the over- or undervaluation of every traded product,” then we would certainly have to exclude all previously existing series for various reasons, including several of the reasons mentioned in the abstract and introduction of this paper. Second, as argued above, theoretically, there is no reason to limit ourselves just to traded products when we talk of competitiveness. Even if labor is not mobile or traded, wages are an important component of competitiveness.

Third, there are at least four basic points to make in defense of the Balassa-Samuelson adjusted WARP index: (1) Whether you like Balassa-Samuelson residuals or not, it is a fact that they are commonly used as measures of under- or overvaluation by academic economists, including Frankel (2006), Rodrik (2008), Subramanian *et al.* (2010), Stahler and Subramanian

(2014), Nicita (2013), Gluzmann *et al.* (2012, 2013) and Chinn *et al.* (2010). The contribution in this paper is how to put BS residuals into a trade-weighted index rather than whether they should be used or not. It should at least be acknowledged that a critique of the index on the grounds that BS residuals are an *a priori* invalid proxy for competitiveness amounts to a critique of a large swath of the academic literature (including the Obstfeld-Rogoff textbook) rather than being a specific critique of this paper.

(2) Given that even I acknowledge that it is not perfect is part of the reason I also introduced the (theoretically and empirically similar) Weighted-Average Relative Unit Labor Cost index. Thus, even those critics who believe that the likes of Paul Samuelson, Bela Balassa, and contemporary writers such as Feenstra, Rodrik, Chinn and Subramanian were all fools to look at Balassa-Samuelson residuals could take solace that I've offered a concrete alternative to Balassa-Samuelson residuals.

(3) As mentioned in the paper, Feenstra, Inklaar, and Timmer (2013) put the empirical Balassa-Samuelson effect on more solid ground when they showed that the Balassa-Samuelson coefficient is remarkably stable over time when you look at the sample of countries that had ICP benchmarks.

(4) Many of the aforementioned papers, including Rodrik (2008), find that Balassa-Samuelson residuals are correlated with economic outcomes. In addition, as I now show in the paper, there is, in fact, a rather tight correlation between GDP per capita and price levels that is hard to deny. Given this, and given the findings of Feenstra *et al.* (2013), it is not clear on what grounds one would want to argue that Balassa-Samuelson residuals are an invalid measure.

This is especially the case in a bilateral setting, where only a selection of the full range of products priced in the ICP (and hence part of PWT) will be relevant. The same criticism holds for the WARULC series: it measures the relative level of unit labor costs, but that does nothing to directly address the index number problem that was used to motivate the analysis. It might be a sensible measure of relative ULC (and it is no doubt the product of substantial data work), but given the motivation at the start of the paper to derive 'theoretically-appropriate measures' it is not clear to me in what way this series is a good solution to the motivating problem.

Response: To be sure, the WARULC series also corrects the index number problem. I may not have made myself clear here, though I did write that "Thus, correcting the 'trading partner substitution bias' problem and the problems arising from country-specific deflators, and including China are more persuasive reasons to prefer WARULC to the IMF's RULC index..." I also hinted at this in both the abstract and introduction, although I didn't explicitly say so until later. In the current draft, I am more clear – and so this was a helpful comment.

Also, how do the BS-WARP and WARULC series relate? Are these equally preferable from a theoretical point of view? Should they be similar?

Response: These are very good questions. I discussed these issues in detail in a previous draft of the paper, and then had to cut this discussion due to space limitations. This discussion is now included in the Additional Appendix.

Furthermore, true like-with-like comparison (or something close to it) is feasible, as demonstrated by Houseman *et al.* (2011) and Inklaar (2012), but relies on detailed price

comparisons to determine how much cheaper products from lower-income countries are and how much of a bias this leads to.

Response: Once again, it isn't theoretically clear why we want to exclude non-traded goods. The ICP and Penn World Tables is actually designed to try and compare like-with-like. Although of course it isn't perfect, this should be seen as an improvement over RER indices which use country-specific deflators, such as the Fed or IMF's CPI-based indices.

Another issue that is not clear to me is what the appropriate weighting should be. Section 3 discusses the Federal Reserve's approach to using export and import weights and discusses criticisms and alternatives. Yet what is absent in that discussion is the question what would be theoretically appropriate. For instance, in the Houseman et al. work, import shares are appropriate since the target concept is an input index. Conversely, for export competitiveness, export shares might be suitable. But for what target concept would a mix of both shares be appropriate? It might not matter a great deal empirically, as suggested on page 12, but again, if the goal is 'theoretically appropriate', then argue what that should be and from perspective.

Response: This is a fair point. I didn't frame the discussion of weights in terms of what the optimal weighting should be. The reason why you might want to combine import, export, and third country weights would be if you wanted to know the impact of relative price changes on the tradables sector as a whole (not just the impact due to export exposure or import exposure). A truly optimal weighting scheme would require detailed knowledge of domestic shipments for each country, which are not readily available for many countries for the full span of time (and even are not straightforward to compute for the US for the entire tradables sector). In addition, as I mentioned, a key finding in this literature is that large theoretical improvements in the weights tend to lead to only small changes in the weights themselves and thus small changes in the indices, and in this paper I confirmed this finding. The exception to this is that time-varying weights do make a difference relative to fixed weights. Thus, given a herculean data task on one hand combined with a seemingly minimal possible payoff makes this seem not worth it (the data work I've done for this paper already took 10 months). For example, the improvements I made already in the weights made a minimal difference, and switching to value-added based weights (which have less coverage) seemingly do not make any real difference to the weights – even to China's trade weight. Using more disaggregated sectoral data to compute the third country weights also seem not to make a large difference.

Finally the empirical implementation, in particular the regression analysis, but also the many graphs. I agree with the author that “theoretical and intuitive concerns should dictate the choice of [an index]” (p29). But as stated in the introduction, the goal is also to have a real exchange rate measure that is informative for changes in trade. The regression analysis in Table 1 suggests that there might be some mileage there, but it feels very exploratory: why only look at imports and not also at exports?

Response: Note that the “import share of total trade” includes exports in the denominator. Thus, this measure is roughly equivalent to looking at the trade balance scaled by how much trading you are doing. I fully agree that it wouldn't make sense to just look at imports. I've tried to make

this more clear in the next draft (as I believe this referee can be forgiven in that my last draft was not clear on this point).

Is there a more explicit link with a gravity model or theoretical foundation?

Response: See what was formerly footnote 17 (now footnote 35), in which I provide an explicit theoretical foundation.

Specifically: is there an expected value for the coefficient on the real exchange rate? Does it inform theoretical parameters? And does the relationship work in out-of-sample prediction (helpful for the applied researcher)? Why only estimate this relationship on US data if the next section provides an international extension with much more data? Why only for the WARULC and not for the BS-WARP?

Response: These are all very good questions. On the first question – estimates from the literature (and from theory) are all over the board for the coefficient on the RER. Theory doesn't even limit us to a certain sign given the potential for third-factor causality (*i.e.*, the Asian Financial Crisis reduced both exports and RERs for many Asian countries). In the new draft I do an explicit out-of-sample test. I also now test internationally as well, and now I test for all new indices.

Reviewer #2: Referee report for "Through the Looking Glass: A WARPed View of Real Exchange Rate History"

This paper aims to extend the concept of a REER to account for differences in price levels and productivities across countries. Author builds on recent work by Fahle, Marquez and Thomas (2008, 2009), who incorporated price level differences into REER indexes and showed that such differences, when combined with increasing global trade share of low income countries (in particular, China), can significantly alter REER index dynamics. Fahle et al. labeled their new index Weighted Average Relative Price (WARP) index. The current paper's contribution is two-fold. First, it extends the original WARP index to historical data for the USA as well as to other countries. Second, the paper proposes a Balassa-Samuelson productivity adjustment for the WARP index to account for the intuitive notion that not all price level differences affect competitiveness.

Response: The other big part of this paper is that I introduce a WARULC index, the unit-labor-cost analogue to WARP.

The paper is well written and I read it with great interest, as earlier work by Fahle et al. is indeed empirically very interesting, but lacking a theoretically-motivated interpretation. Unfortunately, I did not find this key shortcoming of the WARP index addressed in the current paper.

Response: I do not see the WARP index as having any "shortcomings" necessarily, but rather, it provides different information than a Balassa-Samuelson adjusted index, which corrects for

productivity differentials. The reason this is an important correction is that countries such as Bangladesh, which have low wages, may not necessarily be a competitive threat if productivity is low. I do not see this as being a subtle point in need of a theoretical model, but in deference to the referee, I have provided the model in the latest draft (more or less copying a textbook model).

Conventional REER

More specifically, the conventional widely used REER index has well defined theoretical foundations. It measures how changes in international relative prices affect demand for a country's output or some sub-component(s) of the output (exports, domestic competition to imports, trade balance, etc). Theoretical motivation for the REER index comes from a standard Armington demand system: given trade structure and aggregate incomes, how do changes in international relative prices affect demand for a country's output? Loretan's (2005) presentation of the FED's REER index, Bayoumi et al. (2006) construction of the IMF's new REER index, as well as similar reference papers at the EC and BIS all rely on these theoretical foundations of the Armington demand system. They all reference IMF Working Papers from 70s-80s, where the framework was developed. Most importantly, because there is a well-defined theoretical framework underlying the REER index, we have a clear definition for "competitiveness", i.e. price-induced changes in demand for output.

Response: Thank you for this very helpful comment. However, in an Armington demand system, it is relative prices that matter, not changes in relative prices. If a country with low relative prices (let's call it China) suddenly opens up to free trade, then this will lower the average price of imports, and of the average value of goods in the world, even if the price of Chinese goods themselves do not change! It's also worth mentioning that neither Loretan (2005) or Bayoumi (2006) explicitly make reference to an Armington demand system, and Bayoumi uses fixed trade weights – which is also quite clearly not implied by an Armington demand system. (Note – this is not a critique of these papers – it is not clear to me that these papers needed to solve an Armington demand system). In any case, in the latest draft of my paper, I include an Armington demand system, and show that changes in trade volumes arising from changes in costs and country sizes can indeed affect the average price indices, even if no prices themselves change. In the current draft, I also provide a definition of competitiveness (I used the standard dictionary definition).

WARP index

Fahle et al. WARP index points out some intuitively appealing shortcomings of the conventional REER. However, it lacks a clearly spelled out multi-country theoretical framework that would allow one to understand what exactly WARP measures. What is the definition of competitiveness that underlies the WARP index? Or, to put it in model terms, what is the theoretically-prescribed aggregation problem that the WARP index solves?

Response: This is an interesting question. The answer is that in any model in which relative price levels matter, one would prefer a Weighted-Average Relative (WAR) index to an index-of-indices. This is a finding at the core of the central teaching of economics, that “prices matter”.

The more interesting challenge would be to try to write down a model in which only changes in prices matter, but that actual price levels do not. It would likely be difficult to do, but I think one would want to start off by assuming that relative price levels are not known, or cannot be measured. This was the reason Divisia indices have been used for CPI-based RER indices – CPIs are indexes in which there are no internationally comparable levels.

Applying the definition of "competitiveness" from the conventional REER does not work for the WARP index. This is the case because REER index effectively solves a demand side price aggregation problem, while the WARP index introduces supply side considerations (i.e., productivity).

Response: Not necessarily. You can get WARP from a pure demand system, as seen in the model in the appendix of the current draft of the paper. A Divisia index does not solve any known aggregation problem.

With the WARP index productivity affects both prices and income, so it no longer seems reasonable to impose constant aggregate income assumption of the REER framework. Allowing for endogenous income changes, in turn, will affect the aggregation problem. But, as author correctly points out on page 12, WARP index is sensitive to the assumed aggregation weights. So, it is essential to solve the aggregation problem before the WARP index can be constructed.

Response: Actually, there is no assumption of constant aggregate income here. Allowing aggregate income to change will not alter the superiority of WARP in an Armington world (or in a Melitz world). I now present such a model in the appendix.

In sum, my main problem with the paper is that it proposes an index for measuring competitiveness without providing a clear definition for competitiveness. This is not the case for the conventional REER index. Similar to REER, WARP should define and solve an aggregation problem. Instead, author uses theoretically-unmotivated weighting matrix, which we know can significantly affect the WARP index.

Response: In the latest version, I provide a theoretical model and a tight definition of competitiveness. Note, as mentioned above, that other key papers in this literature, such as Loretan (2005) and Bayoumi (2006) do not explicitly make reference to an Armington demand system, and Bayoumi uses fixed trade weights – which is also quite clearly not implied by an Armington demand system and certainly does not arise from any tight definition of competitiveness. I do make what I believe are clear improvements in the trade weights, although I do find that these improvements – as others have found in this literature – do not make a large difference in the indices. The problem with theoretically appropriate weights, as mentioned in the paper, is that the data for these weights do not exist.

Other comments:

1. If paper's main contributions, as stated in the abstract, is to enhance the WARP with a B-S productivity adjustment, then in Table 1 the benchmark for predicting import share should be the

Fahle et al. WARP index, rather than the conventional IMF index. Without WARP index as a benchmark, the contribution of the papers proposed extension of the WARP index cannot be assessed.

Response: This was a very helpful suggestion. In the updated draft I also include the WARP index in Table I. At first I did not test P-WARP (previously called BS-WARP) because P-WARP and WARP yield such similar indices, differing mainly just with the levels.

2. The discussion of trade weights in Section 3.2 should explain why the particular weighting structure is used, i.e., REER weights are theoretically motivated. Because empirically weights have little impact on the REER index, Loretan (2005) simplifies the weighting matrix. However, this empirical justification should not be extended to the WARP index.

Response: Another good suggestion. In the current draft of the paper I discuss weights in more detail, and why I think it will be difficult to make additional substantive improvements in the weights.

3. The title "Theory" for section 5.1 is misleading. The section instead deals with the implementation of price indexes, rather than theoretical underpinnings of price indexes.

Response: Agreed. Fixed.

Reports from First Submission, submitted 4/2014:

Report on ‘Through the Looking Glass’

The authors of this paper argue that the standard way in which statistical agencies construct real exchange rate (RER) series does not provide informative results and in this they follow the arguments of Fahle et al. (2008). Specifically, existing RER series merely track changes over time, instead of providing a comparative level.

Response: This is correct, but that existing RER series do not provide a comparative level is less serious issue than the “trading partner substitution bias” issue. This happens when trade shifts to poor countries, such as China, that have systematically lower price levels.

In moving beyond those arguments, they propose and construct two further RER series. These not only take into account differences in the level of RERs across countries, but in their first additional series they propose to adjust these level differences for Balassa-Samuelson-type differences and in their second series they aim to capture differences in the level of manufacturing unit labor costs.

I have two main questions after reading this paper:

1. How do your RER measures relate to the basic concept you aim to capture, and

Response: In this draft I go into more detail why we should want RER indices which reflect the growing role of trade with poor countries China, as readers may be unfamiliar with Fahle *et al.* (2008), which is actually the seminal paper in this literature and should be required reading for graduate students in this field.

2. What is the criterion for the empirical success of a particular RER measure The current paper is unclear about these questions, while I would argue that these should be the organising principles of a paper such as this.

Response: In this draft I reorganized my paper to emphasize that my view is that one should choose a particular RER index based on theoretical or intuitive considerations rather than testing which RER indices do the best job explaining trade flows or manufacturing employment. Thus, I would choose WARULC over the IMF's RULC index because the former includes China, multiple benchmarks, and time-varying trade weights, and not because it does a much better job explaining the US import share of trade and appears to do a very good job explaining the collapse of relatively more import-competing manufacturing sectors in the US.

Concept

In open macroeconomy models, the real exchange rate serves as a relative price of domestic products relative to foreign products. This price is of clear interest, because if domestic products become cheaper, foreign demand will rise and conversely, domestic demand for foreign products will decline. Translating this simple idea to a real-world setting raises at least two problems, though:

- *What is the scope of 'products' and their 'price'?* Does this refer to the set of products (goods and services) that a country exports? Or the set that it could potentially export if, for instance, a (nominal) depreciation made domestic products more attractive? Presumably, foreign demand will be a function of prices of the (potential) export bundle, but domestic demand for foreign products will in turn be a function of the (potential) import bundle. Is it sensible to try and capture both sets of products in a single index? Presumably only if separate indexes show the same pattern over time. When selecting the scope of products, this also leads to a choice on the appropriate price concept. In introducing the WARP index, the authors suggest using relative price measures that (predominantly) refer to prices of goods for final consumption and investment from PWT. However, what would seem to matter for foreign demand are the prices of exported goods and of imported goods. Following the logic of the Balassa-Samuelson (BS) effect, price differences for such traded products will be smaller than final consumption and investment products. Moreover, Feenstra and Romalis (QJE, 2014) show that there are substantial differences between the relative prices of exported goods and of imported goods. So that would certainly argue in favour of a separate relative export and relative import price index.

Response: In this draft, I address the possibility of using export prices relative to import prices, which are also called the terms of trade. There are, in fact, good reasons why researchers generally **do not use** the terms of trade to measure exchange rate shocks. The first is that, if the dollar appreciates, this could lead to declines in dollar-denominated import and export prices, leaving the terms of trade unchanged even while affecting the trade balance. Generally, terms of trade changes are not very well correlated with other measures of the RER or of trade balances.

Thirdly, for the US, movements in the terms of trade are generally caused by movements in commodity prices including oil, while many researchers are actually more interested in other tradable sectors such as manufacturing. Additionally, tradables' sectors include many non-traded inputs such as land, rent, utilities, and non-traded services, and so there is nothing inherently wrong about using, say, the Federal Reserve's Broad Trade-Weighted Real Exchange Rate index as a measure of competitiveness aside from the shortcomings described in the paper.

Furthermore, the availability of such price data calls into question the need for and accuracy of the BS-WARP index: why adjust the prices of final consumption and investment products using a single cross-country BS coefficient and would that bring the final results closer to export and import prices?

Response: The reason to adjust the prices by productivity is that some countries with low price levels, such as Bangladesh, also have very low productivity and so are not a competitive threat. In this draft, I explain the Balassa-Samuelson logic in more detail. And, to answer the question, none of the other RER measures – from the Fed, the IMF, Fahle et al. or my indices bear any resemblance to the terms of trade. Once again, there is scant rationale to only look at import and export prices when we think about competitiveness.

Seen from this perspective, it is also unclear to me why a relative unit labor cost measure would be a helpful measure: **what matters for a foreign buyer is the price, not the price relative to wages.**

Response: The reason that wages matter for competitiveness is that labor is a large non-traded and largely immobile input into the production process. Thus, if you are a firm looking whether or not to locate a factory in the US or China, it stands to reason that two factors you would want to look at are: (1) the going wage of workers, and (2) how productive these workers are. In this draft, I go into more detail on this point, and also mention that in most standard trade models, including Heckscher-Ohlin, the Melitz Model, and in Ricardian Models, wages and other input costs clearly matter.

Of course, for analysing why the price of a product is relatively high, it can be sensible to look at input costs, but that would seem to ask a different question.

Response: Once again, if workers were just as productive in the US as in China, but wages were lower in China, why wouldn't a profit maximizing firm relocate to China? Note that this firm wouldn't necessarily need to lower its prices as it lowers its costs. The implicit assumption I am making is that firms respond to incentives/prices.

- *What is the scope of 'foreign'?* Does this refer to the set of countries buying a country's exports or the countries from which imports are sourced? Or again, is the set of potential trading partners larger than the actual set (and how would we determine what this set is) and should these two sets be combined into a single index? I would immediately agree that the FRB series (or any other official series) is not very explicit or methodologically pure on any of these issues, but a serious academic paper on this topic should be.

Response: Interesting philosophical point. In this case, I decided to be conventional and not to innovate much on this front and just stuck with the method the BIS has done, which is very similar to the methodology of the Federal Reserve. I also try using GDP weights, and get generally similar results, although I tend to find that US prices are higher than with trade weights, since the US trades more with rich countries. Intuitively, if there is no observed trade between two countries, then this is probably an indication that the trade costs are relatively high, and that a RER appreciation vs. this country is not likely to impact competitiveness very much. This is an argument for using trade weights rather than GDP weights. In the current draft, I have expanded the number of trading partners from 40 to 153 – the maximum possible given data limitations.

Empirical criterion

Given indexes that capture an appropriate concept, the question arises how to establish whether this represents a price that is relevant for foreign buyers of domestic products and domestic buyers of foreign products. It is not at all clear to me what the ‘intuitive priors’ are that the authors refer to on p29 when choosing the ‘share of manufactured imports of total manufacturing trade’ as the dependent variable and the subsequent model specification is very ad-hoc.

Response: The “intuitive priors” are that, as GDP increases, imports increase as per the gravity equation, and as foreign GDP increases, exports increase. As foreign prices increase relative to domestic prices, agents can make more money by selling in the foreign market. In this draft I explain this in more detail, and have also introduced a simple model.

Given the extensive literature estimating export and import demand, it should certainly be possible to formulate a model with sounder theoretical foundations.

Response: I have introduced a model in the current draft, albeit in the footnotes given the space constraints.

Review of “Through the Looking Glass: A WARPed View of Real Exchange Rate History”, MS # 19432

This paper computes alternative measures of Real Effective Exchange Rates (REER) that propose to improve on existing indices in six dimensions. First, it relinquishes using indices, and reserves information about relative price levels. Second, it computes directly the (trade) weights used to average bilateral exchange rates, lets them vary over time, and introduces variations between arithmetic vs. geometric averages of bilateral exchange rates. Third, it corrects for Balassa-Samuelson effects. Fourth, it performs an aggregation of relative unit labor costs, which once again preserves information on their relative levels, and once again aggregates bilateral comparisons using observed (and time-varying) trade weights. Fifth, it adjusts these measures for “domestic competition”. Sixth it extends (some of) these corrections to a dozen other countries than the US. The corrected measures display dynamics that are quite different from the conventional ones.

The paper goes through a formidable data gathering effort, and the authors should be commended for this. The data they put together could eventually be of use to the profession,

though in a way that is not obvious at this stage. My main comment about the paper is that it lacks focus, an analytical framework in which to think about the adjustments it implements, and perhaps most importantly, a question. The fact that different averages, using different data, imply different conclusions is not particularly enlightening nor surprising.

Response: First of all, thank you very much for your report as this is very helpful. In order to increase the focus, in this draft I eliminated several whole sections of the paper, on the time-varying BS-WARP index and also on the indices adjusted for domestic completion (now in the Additional Appendices). Secondly, I add in a simple analytical model to motivate the Balassa-Samuelson adjustment. Thirdly, in this draft I am more explicit that the goal of this paper is merely to produce RER indices that applied researchers can take “off-the-shelf” that do not suffer from the same clear “trading partner substitution bias” problem as do existing publicly-available RER indices, or from the omission of China as the IMF’s RULC index does. In this draft I am also more clear that, at least for the US, all three of the class of “Weighted Average Relative” (WAR) indices in fact imply similar conclusions about US competitiveness over the past 40 years. In this draft I also discuss in more detail the theoretical motivation for creating indices which adequately incorporate the rise of China.

Before going through my comments in more details, let me start by a general point about the measurement of REER. Tracking their level over time requires two elements: (i) a measure of relative price levels, which of course the standard price indices (CPI, PPI, WPI, GDP deflator) do not provide, (ii) a time-varying measure of the composition of price indices in both countries. Both of these objects are exceedingly hard to obtain, which is the main reason why most REER are only meaningful in growth rates. Tracking the dynamics of REER over time requires two additional elements: (iii) changes in trade weights, and (iv) changes in country coverage (equivalently, I think).

Response: Note that the PWT actually provides annual estimates of relative price levels, making (i) and (ii) admittedly not that difficult to attain in the case of WARP, but much more difficult in the case of Unit Labor Cost-based measures. Also note that the re

The paper under review attempts to improve on all these aspects. As a result, it lacks focus, as well as, I would argue, an important question.

Response: In this draft I cut out several sections of the paper to improve the focus. I, of course, strongly disagree with the notion that measuring the real exchange rate properly is not an important research question. There is a good case to be made that the real exchange rate is second only to the real interest rate in terms of the most important prices in an economy. In addition, there is a growing amount of evidence that real exchange rate movements have played a key role in the sudden collapse of US manufacturing (Campbell, 2014), the Asian Financial Crisis (see Chinn, 2000) and also the problems in Japan since 1992 and the problems in several European economies during the past 6 years. While I would concede that indexing and measurement issues do not seem like sexy topics and the issues are a bit subtle, it is easy to make a case that these are, in fact, critically important issues.

In my view, a measure of REER that would embed information on (i) and (ii) would already be an important contribution. Unfortunately, the paper under review never clearly explains how price levels are measured across countries. It refers to “PPP data via the Penn World Tables for every several years between 1990 and 2010” (page 10), simply “PPPt” in equation 3.7, or “manufacturing PPP using PWT v8.0 methodology” (page 21).

Response: “PPP data via the Penn World Tables” refers to the variable labeled “PPP” in the Penn World Tables up to version 7.1. In the current version, however, only the “price level” is given, which is PPP divided by the nominal exchange rate. In the current draft I am more explicit about this.

I have no idea what these sentences mean. As far as I know, the only source of information on price levels across countries comes from the International Comparison Program (ICP), which underpins the PPP computations in the Penn World Tables. Since no explanations are given in the paper under review as regards data sources, I have to believe this is what is used. But these are collected every five years at best, and suffer from the usual arbitrariness inherent in survey based data collection. So I am afraid I am really not sure how the paper under review is addressing point (i). Point (ii) is never mentioned, so I don’t think it is addressed either. So the paper cannot be about points (i) or (ii), i.e., about the levels in international prices. Still, it claims in many places that it is.

Response: The Penn World Tables provides PPP data annually – in this draft I mention so explicitly. The nominal exchange rate and PPP levels combined give you the price level between countries. The PWT does this based on the ICP for the base years, and as I write in the paper, they extrapolate in the intervening years using country-specific deflators.

Point (iii) is made a big deal of.

Response: I certainly did not intend to make a big deal of the trade weights in the improvement of WARP. I proposed a very sensible, albeit minor correction in the direction of plausibility and found it had a minor impact on both the trade weights and the indices. I spend less than two pages on this issue, and did not believe it merited reference it in the abstract or conclusion. For the IMF’s RULC index, the IMF’s trade weights were fixed, outdated, and did not include China, and so changing the trade weights was important, but a relatively straightforward task.

It makes sense that the trade weights used in computing the changes in REER should vary over time. But computing them does not go without problems. Equations (3.2) and (3.3) report intuitive definitions of trade weights, which should be the ones used to compute them. But the volumes of imports and exports are objects that are difficult to observe: it is typically their values that are collected in trade data. So computing the trade weights in equations (3.2) and (3.3) requires data on import and export prices, which are far from readily available. (This is in fact one of the most important improvement Feenstra et al have brought to PWT 8.0). And even if prices were available, the issue of pricing to market would make the computation of trade weights problematic. Without a theory, it is difficult to know whether an observed year-to-year increase in real trade shares should in fact be reflected in a REER: what if it is just reflecting local currency pricing?

Response: In fact, equations (3.2) and (3.3) refer to trade values rather than volumes. If we added up volumes of different goods, including, literally, apples and oranges, we would need to weight them somehow. Prices are a natural weight (much better than actual weight), and this is why all trade-weighted RER indices I know of use trade values. Using “real” trade weights might be an interesting exercise, but I do not have access to this data, and my intuition is that this adjustment would not make a large difference on the trade weights or the overall indexes.

These issues may be the reason why the Fed is using the (somewhat arbitrary) approach summarized on page 11.

Response: Perhaps, but this is very unlikely. As the creators of the Fed’s index itself wrote, they found that more complicated trade-weighting methodologies in the direction of plausibility tended to have a minor impact on the overall impact, and thus kept it simple. The method I propose is only slightly more complicated, and has been used by the BIS and others.

But then at the end of this all, we are told at the bottom of page 11 that the results are “little-changed” when alternative trade weights are used that reflect the actual composition of imports and exports. And yet, we are told later in the paper that the key difference in the improved indices “WARULC” comes from the inclusion of China, i.e., something like my point (iv). How can the inclusion of China matter if trade weights do not? Trade weights change with the country coverage. This is highly confusing, and I think it comes from the fact that the paper under review does not attempt to present sequentially the improvements it proposes. As a result, it is impossible to understand which of the 4 points I list above matter, where, and why they do.

Response: Apologies for my lack of clarity. In this draft I clearly state that using the Fed’s trade-weight methodology vs. my methodology (which others have also used) doesn’t have a large affect on the actual trade weights, and thus doesn’t have a large affect on the indexes. If you dropped China from the sample, the trade weights would obviously change a lot, and thus the indices would change as well (as I show). Also, if you used the IMF’s fixed trade weights instead of the Fed’s trade-weights, the trade weights would be quite different and thus so would the indices (I also show this).

Instead, we have some comparisons between indices that are all equally a-theoretical. It is impossible to know what to make of them. By way of contrast, here is a suggestion of something that would constitute an important contribution. Suppose points (i) and (iii) were taken care of, in a way that were transparent both empirically and theoretically. Then one could compare the existing indices, where (i), (ii), and (iii) are not taken care of, with the improved one, and thus infer the effect of changes in the composition of price indices on REER. I am not saying the paper under review should conduct this exercise. But this is an example of a precise and important question.

Response: I agree that this is an example of a precise and important question, and I thank you for the suggestion. However, note that “fixing” the points (i-iv) would not solve the two key major problems in existing indices that this paper is designed to address. These are the “trading partner substitution” bias problem that means that existing indices do not adequately account for the rise

of China and the fact that WARP is not adjusted for productivity differences (which matters according to the Penn Effect).

I find the successive “corrections” implemented in sections 4, 5, and 6 hard to understand. The Balassa-Samuelson effect drives differences in aggregate price levels that are due to non-traded goods. But does that mean a measure of “competitiveness” should control for levels in GDP per capita?

Response: The Balassa-Samuelson effect would clearly seem to say yes. If Bangladesh has a low overall price level, then this does not necessarily mean they are a competitive threat if their productivity is also low. GDP per capita is one measure of productivity, and the empirical Balassa-Samuelson Effect is one of the most robust results in Open-Economy Macroeconomics. Note that Balassa-Samuelson residuals are commonly used to assess RER valuations, including by Rodrick, Subramanian, and Menzie Chinn.

Does that not mean that they should focus on traded goods? (Which incidentally runs into the issue documented by Engel (1999) that traded goods tend to be responsible for the anomalous behavior in the real exchange rate). I am not sure either I understand what it means to adjust for “domestic competition”. I agree that openness depends on country size. But does that mean country size (and openness) should be controlled for when computing a measure of “competitiveness”? Is openness not endogenous to a country’s competitiveness? At the very least I would think this depends on what the measure is destined to achieve. Incidentally, controlling for openness in equation (6.1) means including a measure of the trade balance in the REER. Is that not the main reason why the corrected measure correlates so highly with the current account?

Response: For the purposes of simplicity, I have moved the section of the paper that adjusts for domestic competition to the appendix. The purpose of this is that a country that is mostly closed to trade will not be much hurt by even large exchange rate swings, while a very open economy will be. Controlling for openness is not in fact equivalent to controlling for the trade balance – the US has a large trade deficit and yet has a small overall level of trade, and for the US, the two measures are not clearly correlated over time.