AEA Papers and Proceedings 2019, 109: 444–449
https://doi.org/10.1257/pandp.20191005

Measuring Productivity: Lessons from Tailored Surveys and Productivity Benchmarking†

By David Atkin, Amit K. Khandelwal, and Adam Osman*

Economists have long recognized that the reason some countries are richer than others is not primarily due to differences in their endowments or resources, but because of how effectively their firms deploy those resources to generate economic activity. This prompts a set of obvious questions. How large are productivity differences across firms? What drives this dispersion? What policies are most effective at raising productivity? Answering these questions is an active area of research (see the review by Syverson 2011) and central to this goal is the ability to accurately measure the productivity of firms.

What the researcher typically wants is a measure of physical output conditional on physical inputs, termed quantity-based productivity (TFPQ). This requires data on input and output quantities that are not typically available. In cases where these data are available, quantities are likely measured with substantial error since they cannot be easily read off accounting statements. Even if well measured, product specifications and quality levels can vary dramatically across firms and within firms across product lines—variation that is not well captured by disaggregated product categories in typical administrative datasets. This makes it difficult to both measure productivity for firms that produce many varieties or to compare productivity across firms making different varieties. Multi-product firms pose further challenges since output and input mixes vary even more widely across products than within.

As most firm-level datasets only provide expenditure and revenue data, much of the literature relies on revenue-based productivity (TFPR) measures that also capture differences in markups and quality across firms. However, if a firm's capabilities come from its ability to produce both quality and quantity, TFPR may be closer to the object of interest even though it confounds forces unrelated to productivity.

The existing literature has pursued various approaches to understand and to mitigate these measurement issues by drawing upon additional data (e.g., prices) alongside (often strong) identifying assumptions.[1]

We take a different approach. We develop tailored surveys focusing on a specific industry—flat-weave rugs—that directly address many of these measurement issues through the combination of detailed product specifications and external assessments of quality. The surveys allow us to calculate not only quantity productivity (the ability to produce quantity with a given set of inputs), but also quality productivity (the ability to produce quality with a given set of inputs) and capabilities (the combination of the two, essentially a TFPQ measure using quality-adjusted quantities).[2]

To better understand the shortcomings of standard productivity measures and potential remedies, we compare survey-based productivity measures to productivity benchmarking exercises

*Atkin: Department of Economics, MIT, The Morris and Sophie Chang Building, 50 Memorial Drive, E52-550, Cambridge, MA 02142 (email: atkin@mit.edu); Khandelwal: Columbia Business School, Uris Hall 606, 3022 Broadway, New York, NY 10027 (email: ak2796@columbia.edu); Osman: University of Illinois at Urbana-Champaign, 109 David Kinley Hall, 1407 W. Gregory, Urbana, IL 61801 (email: aosman@illinois.edu). We thank AbdelRahman Nagy and the Egypt field team. We acknowledge generous funding from the International Growth Centre, Private Enterprise Development for Low-Income Countries, Innovations for Poverty Action, Economic Growth Center at Yale University, McMillan Center at Yale University, and the Jerome A. Chazen Institute for Global Business at Columbia Business School.

†Go to https://doi.org/10.1257/pandp.20191005 to visit the article page for additional materials and author disclosure statement(s).

[1] See De Loecker and Goldberg (2014) for a review of identification assumptions and measurement issues in production function estimation.

[2] Hallak and Sivadasan (2013) also explore multidimensional firm productivity.

which we argue are closest to true productivity. We find that standard TFPQ performs poorly at measuring quantity productivity, shows excessive dispersion across firms, and is inversely correlated with quality productivity. Controlling for product specifications—effectively making apples-to-apples comparisons—goes a long way toward remedying those deficiencies. Although TFPR does better than TFPQ at capturing broad capabilities, it performs worse than methods that combine survey information with explicit quality measures.

I. Survey Design and Data

Our data come from surveys we designed and administered on 219 rug-making firms in Fowa, Egypt. These firms produce a type of kilim rug called "duble" using double-treddle foot powered looms. As part of a randomized experiment exploring the impact of exporting we recruited all firms with one to five workers making this type of rug.

Rug producers receive orders with a particular set of specifications that include the design, thread types, and thread count. Producers prepare the appropriate inputs, install the threads on the loom, and weave the rug. Although duble rugs are already a subset of a ten-digit Harmonized System (HS) product code—the finest product classification in trade data—there are many varieties (we observe 435 unique combinations of specifications).

In addition to rugs having different specifications, rugs also differ in quality. Unlike specifications—codifiable attributes of the rug that are typically chosen by the buyer—quality depends on weaving technique and is difficult to codify or contract on. For example, how flat the rug lies is determined by how skillfully the firm installs the thread on the loom, and whether the threads are held correctly while weaving.

We created a survey instrument to address the measurement issues noted above in contexts where output varies in quality and firms produce many varieties. We administered six rounds of surveys at the product-line level capturing the rug produced in the prior month. (As production runs last longer than a month in this industry, this was almost always a single variety of rug.) These surveys recorded detailed rug specifications; prices and quantities of all inputs and outputs; and labor hours spent on production and preparation activities. We also hired an independent quality assessor who graded each rug that the firm was working on at the time of the survey across 11 different dimensions (grading on a 1 to 5 scale).[3]

Additionally, we set up a controlled laboratory in a rented space where all firms were paid a flat fee for their head weaver to produce a 0.98m^2 rug with identical specifications using identical material inputs and capital equipment we provided. We recorded dimensions, weight, and time taken to weave the resulting rug, and sent the rugs to be scored anonymously by both our quality assessor and a local professor of handicraft science. Atkin, Khandelwal, and Osman (2017) provide further details on the sample, rug production, and the laboratory.

Online Appendix Table 1 provides summary statistics of the survey and lab, and online Appendix Table 2 shows that our six product specifications (thread type, thread count, design difficulty, number of colors, market segment, duble subcategory) capture rug varieties relatively well—specifications explain about half the variation in prices, output and revenue, and dimensions such as thread count and type have the expected signs.

II. Measuring Productivity

We calculate several productivity measures from the survey data. The first measure we call "unadjusted" productivity because, as in the existing literature, it does not adjust for the fact that different firms produce rugs with different specifications (i.e., different varieties). We estimate unadjusted TFPQ (ϕ_u) from a Cobb-Douglas production function

$$x = \phi_u l^{\alpha_l} k^{\alpha_k} e^{\epsilon}, \tag{1}$$

where x is output in square meters, l is labor hours, k is capital (number of looms), and ϵ is measurement error.[4] For transparency, we estimate (1) in logs over every firm-round observation using OLS and recover

[3] The dimensions are: corners, waviness, packedness, weight, touch, warp thread tightness, firmness, design accuracy, warp thread packedness, inputs, and loom.

[4] At this level of disaggregation, the production function is best characterized as Leontief in materials. The unit of analysis is the firm-round level.

ϕ_u by exponentiating the residual. The online Appendix replicates the analysis estimating (1) using a control function.

Although this formulation is standard, a number of features reduce measurement concerns compared to other settings. First, we observe quantities of x, l, and k rather than revenues and expenditures. Second, given the simple technology there are essentially no other inputs used in production (e.g., no accounting, logistics, human resources). Third, we recorded the inputs used for each specific rug produced so there is no error in allocating inputs to outputs.

Our second measure, "specification-adjusted" productivity, is more novel because it controls for differences in the variety mix across firms that may make standard TFPQ measures misleading. To guide our specification adjustment, we place more structure on equation (1) by assuming $\phi_u = \phi_a e^{\lambda\gamma}$, where λ denotes the vector of specifications which affect how quickly a rug can be produced (e.g., a high thread count rug requires more labor and capital inputs) and γ are parameters to estimate. The term ϕ_a is specification-adjusted TFPQ that is recovered from estimating the production function conditioning on the six specification controls.

These two measures essentially capture how many labor hours firms require to produce rug quantity, potentially controlling for the specifications of the rug; we call these *quantity* productivity. While the literature typically explores a single dimension of productivity, as discussed above, similar specification rugs also vary substantially in quality. Thus, there is a second dimension of productivity that also raises revenues:[5] the skill of a firm at producing quality from a given set of inputs. We term this *quality productivity*, or TFPZ.

It is necessary to construct a quality index in such a way that quality and quantity productivity estimates can be compared and aggregated. To do so, we let the consumers' relative valuation of quantity and quality guide us. For simplicity, we make the assumption that consumers have CES demands between rugs and another good y, where consumers trade off the quality and quantity of rugs as follows: $U = \left(\left(\Pi_j q_j^{\theta_j} x\right)^{\frac{\sigma-1}{\sigma}} + \left(y\right)^{\frac{\sigma-1}{\sigma}}\right)^{\frac{\sigma}{\sigma-1}}$. The vector θ determines the trade-off between quantity and the 11 dimensions of quality $\mathbf{q}$ indexed by j. This implies demands

$$\ln x = (\sigma - 1)\sum_j \theta_j \ln q_j - \sigma \ln p + c, \tag{2}$$

where p is the price of the variety and c is a function of total expenditure and the rug price index that is common across varieties. We recover the θ_js by regressing $\left(\ln x + \sigma \ln p\right)/\left(\sigma - 1\right)$ on our 11 quality metrics and setting $\sigma = 2.74$ based on Broda and Weinstein (2006).[6]

We next conjecture a production function for producing consumers' valued quality, $\Pi_j q_j^{\theta_j}$, with the same functional form as the quantity production function

$$\Pi_j q_j^{\theta_j} = \zeta_u l^{\beta_l} k^{\beta_k} e^{\varepsilon}, \tag{3}$$

where ζ_u is the residual after conditioning on labor and capital inputs.[7] We assume $\zeta_u = \zeta_a e^{\lambda\delta}$, with the residual increasing in the firm's quality productivity, ζ_a, and allowing the residual to also depend on specifications (for example, ensuring high quality for a high thread count rug may require more inputs than for a low thread count rug).[8] Thus, as with quantity productivity, we estimate two variants of quality productivity: unadjusted TFPZ (ζ_u) and specification-adjusted TFPZ (ζ_a).

Given our assumptions on supply and demand, it is straightforward to aggregate quantity and quality productivity by forming a production function for the $\Pi_j q_j^{\theta_j} x$ aggregator valued by consumers ($\Pi_j q_j^{\theta_j} x = \zeta_a \phi_a e^{\lambda(\gamma+\delta)} l^{\alpha_l+\beta_l} k^{\alpha_k+\beta_k} e^{\epsilon+\varepsilon}$). The implied productivity aggregator, which we term firm capability or specification-adjusted TFPC, is the product of specification-adjusted TFPZ and TFPQ ($\zeta_a \phi_a$).[9] Similarly, unadjusted TFPC is $\zeta_u \phi_u$.

[5] Atkin, Khandelwal, and Osman (2017) shows that prices conditional on specifications increase in quality.

[6] This is their average elasticity estimate within the six-digit HS category for these rugs (HS 570231).

[7] In principle we could also adjust inputs with measures of input quality, particularly worker skill.

[8] Atkin, Khandelwal, and Osman (2017) implicitly assume $\beta_l = \beta_k = 0$ as skill (rather than l and k) primarily determines quality in this industry. Here, we posit similar production functions for quality and quantity.

[9] This approach mirrors the price index literature with equation (2) acting as a hedonic regression that quality adjusts quantities before estimating the production function.

Finally, we also estimate standard TFPR from equation (1) but replacing x, l, and k with rug revenues, labor expenditures, and the value of the capital stock, respectively.[10]

We compare the survey-based measures with productivity benchmarks from the controlled lab. The lab provides direct measures of quantity productivity in meters squared per unit input: Lab TFPQ $= 0.98\,\mathrm{m}^2/\left(l_{lab}^{\hat{\alpha}_l} k_{lab}^{\hat{\alpha}_k}\right)$, where l_{lab} is the hours taken to produce the rug in the lab, and $k_{lab} = 1$ is the number of looms. We calculate lab quality productivity, Lab TFPZ $= \Pi_j q_{lab,j}^{\hat{\theta}_j}/\left(l_{lab}^{\hat{\beta}_l} k_{lab}^{\hat{\beta}_k}\right)$, by combining the anonymized quality assessments for the lab rugs (averaging over the two experts' grades) with the $\hat{\theta}_j$s from regression (2). The $\hat{\alpha}$s and $\hat{\beta}$s come from the specification-adjusted production function estimates above. Lab capabilities, Lab TFPC, is simply the product of Lab TFPQ and Lab TFPZ.

As we are able to ensure that inputs and product specifications are identical across firms, we believe that the lab measures contain the least measurement error and come closest to reflecting firms' true productivity.[11] Thus, we treat them as benchmarks with which to assess our survey measures.

To summarize, the surveys provide two measures of quantity productivity (ϕ_u, ϕ_a), two of quality productivity (ζ_u, ζ_a), two of capability ($\zeta_u \phi_u, \zeta_a \phi_a$), and TFPR. The controlled lab provides three benchmarks: lab quantity productivity (Lab TFPQ), lab quality productivity (Lab TFPZ), and lab capabilities (Lab TFPC).

For each firm, we have one survey-based measure for each survey round. To reduce noise, we take firm-level averages over all post-baseline rounds and present all the productivity measures relative to the mean.[12] The online Appendix provides further details on implementation and the production function estimates.

III. Comparing Productivity Measures

We now explore the relationship between the various productivity measures and draw conclusions for practitioners working with less-rich datasets. We also discuss the dispersion in productivity across firms implied by each measure, a key moment of interest in the productivity literature.

RESULT 1 (Importance of Adjusting for Product Specifications): Comparing unadjusted TFPQ across firms is challenging when specifications vary substantially.

Figure 1 plots both unadjusted and specification-adjusted TFPQ against Lab TFPQ—the measure we believe is closest to true quantity productivity.[13]

Consistent with the claim above, although the slope is positive ($\beta = 0.13$), unadjusted TFPQ only weakly correlates with Lab TFPQ (corr $= 0.02$). Specification-adjusted TFPQ has a steeper slope and a stronger correlation with Lab TFPQ ($\beta = 0.51$, corr $= 0.14$). This shows the value of finer product-category controls for accurately measuring quantity productivity.

RESULT 2 (Quantity versus Quality Productivity): In this industry, as in many others, consumers place substantial value on quality. Our prior is that firms that are able to produce high quality are highly skilled, and so can also produce products with a given set of specifications more quickly. If there is a strong positive correlation between the two, then quality-productivity measures may do a satisfactory job at capturing a firm's broader capabilities.

Figure 2 shows two plots. The first reveals a strong negative relationship between unadjusted TFPQ and unadjusted TFPZ (black). Thus, in the absence of specification controls, quantity and quality productivity are *negatively* correlated: firms that make lower quality rugs produce more quickly. However, and further

[10] Rug revenues and the value of k come from direct survey questions. Labor expenditures equal wages paid to employees and the take home pay of weaver-owners. Values are adjusted using the monthly CPI.

[11] Since the loom, specifications, and inputs are identical for all firms in the lab, we do not need to specification adjust to compare across firms.

[12] The experiment in Atkin, Khandelwal, and Osman (2017) showed that inducing firms to export raised productivity. To ensure that we are not combining estimates for the same firm pre and post treatment, we only include the post treatment rounds.

[13] The figures show both the line of best fit, the slope and significance of this line, the correlation coefficient, as well as a bin scatter of observations (each dot reflects about ten firms). The online Appendix reports a correlation matrix for the various measures.

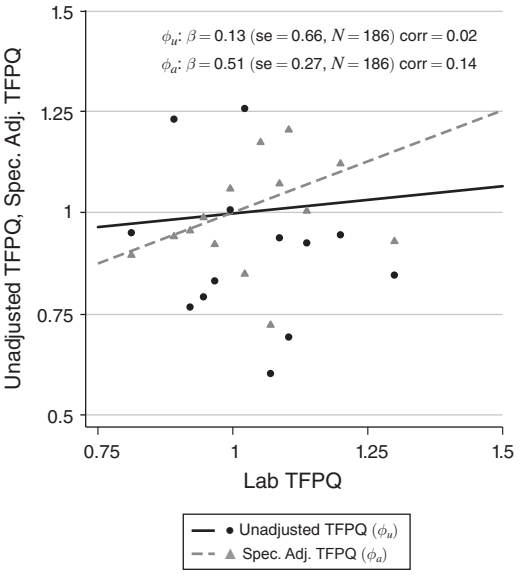

FIGURE 1. ADJUSTING FOR PRODUCT SPECIFICATIONS

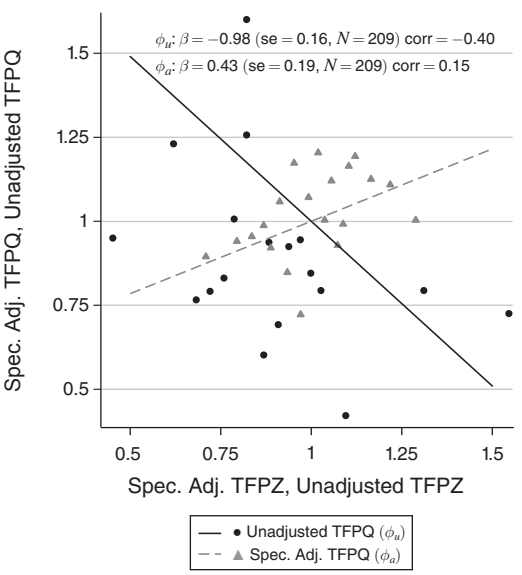

FIGURE 2. QUANTITY VERSUS QUALITY PRODUCTIVITY

showing the importance of adjusting for specifications, this relationship *flips* when we adjust for specifications (the second plot in gray, specification-adjusted TFPQ against specification-adjusted TFPZ). Consistent with our prior, quantity and quality productivity are positively related. More capable firms take longer to manufacture rugs only because they typically make varieties with more demanding specifications.

Thus, and as we show more directly below, in the absence of specification controls, TFPQ may be a misleading measure of broad capabilities given the strong negative correlation between unadjusted quantity and quality productivity.

RESULT 3 (TFPR as a Proxy for TFPC): If capabilities are multidimensional, and consumers value quality, TFPR may be preferable to TFPQ-based measures since higher prices and revenues may capture the ability to produce high quality. Figure 3 explores this claim by comparing several of our productivity measures to Lab TFPC, the capability measure that combines quality and quantity productivity from the lab.

Consistent with the discussion above, unadjusted TFPQ is a misleading measure of capability: it is negatively correlated with Lab TFPC (black diamond). However, TFPR (gray circle) does indeed mitigate this measurement issue since it is positively correlated with Lab TFPC. Although the relationship is weak, this reversal of slope relative to unadjusted TFPQ reveals that TFPR may be a more suitable proxy for a firm's capability than TFPQ if product specifications are unavailable. Specification-adjusted TFPQ (black triangle) is more strongly positively correlated with Lab TFPC. As shown in Result 1, it more accurately captures quantity productivity, and as shown in Result 2, quantity and quality productivity are positively correlated after specification-adjusting. Reassuringly, specification-adjusted TFPC (gray square), which combines specification-adjusted TFPQ and TFPZ, has the strongest positive relationship with Lab TFPC.

RESULT 4 (Unadjusted TFPQ Overstates Dispersion more than Specification-Adjusted TFPQ): Table 1 provides 90–10 ratios for the various productivity measures (we relegate distribution plots to the online Appendix). Dispersion in Lab TFPQ is over three times smaller than unadjusted TFPQ. Adjusting for specifications closes about half this gap. This suggests that dispersion in standard datasets may partially

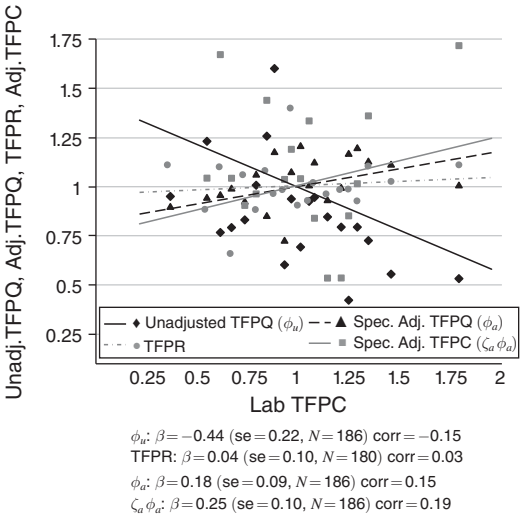

FIGURE 3. TFPR AS SUBSTITUTE FOR TFPC

TABLE 1—PRODUCTIVITY DISPERSION (90–10 RATIOS)

Lab TFPQ	1.3	Lab TFPZ	2.2
Lab TFPC	2.3	TFPR	2.7
Unadj TFPQ (ϕ_u)	4.7	Adj TFPQ (ϕ_a)	3.1
Unadj TFPZ (ζ_u)	2.5	Adj TFPZ (ζ_a)	1.5
Unadj TFPC ($\zeta_u \phi_u$)	4.3	Adj TFPC ($\zeta_a \phi_a$)	3.5

Note: Table reports 90–10 ratios for productivity measures.

reflect product differentiation rather than differences in underlying productivity.

RESULT 5 (TFPZ Dispersion is Large)*:* Table 1 reveals large dispersion in quality productivity. The 90–10 ratio in Lab TFPZ is 2.2. From the surveys, the unadjusted and adjusted TFPZ ratios are 2.5 and 1.5, respectively. This suggests that even within a very narrowly defined product category, there is large quality variation across firms.

RESULT 6 (TFPC is More Dispersed than TFPQ and TFPZ)*:* Capabilities are even more dispersed than either quantity or quality productivity. The 90–10 ratio for Lab TFPC is larger than that for Lab TFPQ and Lab TFPZ (similarly for specification-adjusted TFPC). An implication of the fact that quantity and quality productivity are positively correlated, this result suggests that the broad capabilities of firms may be more dispersed than a single dimension of productivity. To our knowledge, this is the first attempt to document dispersion in capabilities through direct measurement.

IV. Concluding Remarks

We close with a summary of this measurement exercise. First, standard TFPQ performs poorly at measuring quantity productivity. Using product specifications to make apples-to-apples comparisons substantially raises the correlation with the lab benchmarks and halves the gap in measured productivity dispersion between survey and lab measures. Second, firms differ substantially along a second dimension of productivity—their ability to produce high-quality products. Finally, if researchers are interested in broader capabilities of firms, TFPR—for all its imperfections—may be a better proxy than (unadjusted) TFPQ. TFPQ is likely to perform particularly poorly in settings like ours where more capable firms make products with more demanding specifications that take longer to manufacture. But, tailored surveys that collect product specifications and direct measures of quality may be the best path to understand productivity dispersion across firms.

REFERENCES

Atkin, David, Amit K. Khandelwal, and Adam Osman. 2017. "Exporting and Firm Performance: Evidence from a Randomized Experiment." *Quarterly Journal of Economics* 132 (2): 551–615.

Broda, Christian, and David E. Weinstein. 2006. "Globalization and the Gains from Variety." *Quarterly Journal of Economics* 121 (2): 541–85.

De Loecker, Jan, and Pinelopi Koujianou Goldberg. 2014. "Firm Performance in a Global Market." *Annual Review of Economics* 6: 201–227.

Hallak, Juan Carlos, and Jagadeesh Sivadasan. 2013. "Product and Process Productivity: Implications for Quality Choice and Conditional Exporter Premia." *Journal of International Economics* 91 (1): 53–67.

Syverson, Chad. 2011. "What Determines Productivity?" *Journal of Economic Literature* 49 (2): 326–65.