Correspondence

Measuring Power in International Relations

Caleb Pomeroy
Michael Beckley

To the Editors (Caleb Pomeroy writes):

In "The Power of Nations: Measuring What Matters," Michael Beckley persuasively argues that aggregate measures of power systematically exaggerate the capabilities of relatively populous countries.[1] Traditional measures of capabilities—namely, gross domestic product (GDP) and Composite Index of National Capability (CINC) scores—fail to deduct state liabilities, including production, welfare, and security costs. Beckley's proposed solution is to replace measures of aggregate capabilities with a measure of GDP multiplied by GDP per capita (GDP × GDPPC), thus penalizing countries with large populations. The variable exhibits strong theoretical appeal and case-level evidence. For quantitative international relations scholars, however, the most convincing argument for the adoption of this variable hails from the variable's model fit improvements in the "majority of studies published in leading journals over the past five years," as noted in the article's summary. I raise two issues with such a claim.

First, the variable that Beckley proposes does not appear to improve model fits in the majority of replicated studies. He soundly reasons that if the replacement of GDP or CINC scores with his GDP × GDPPC variable improves model fit, then the proposed variable better explains the outcome of interest. As evidence, Beckley compares Akaike information criterion (AIC) scores and tallies improvements based upon GDP versus GDP × GDPPC separately from CINC versus GDP × GDPPC for a given study's replication.[2] This approach to model selection is relatively unorthodox, however. A sounder approach consists of a simultaneous comparison among all three models. Unless the proposed variable outperforms both of the existing variables, then at least one of the existing variables suffices.

Although Beckley rightly points out that models specified with GDP × GDPPC exhibit lower AICs than those that employ CINC scores in seventeen of the twenty-four studies that he examines and GDP in eleven of the twenty-four studies, in only ten of the studies does this measure exhibit an AIC lower than both of the models specified with GDP and CINC scores. Furthermore, these differences must meet some threshold in order to conclude significant fit improvement, typically a difference of

Caleb Pomeroy is a Ph.D. student in the Department of Political Science at Ohio State University. He thanks Robert Bond, Skyler Cranmer, Jared Huling, and an anonymous reviewer for feedback.

Michael Beckley is an associate professor in the Department of Political Science at Tufts University.

[1] Michael Beckley, "The Power of Nations: Measuring What Matters," *International Security*, Vol. 43, No. 2 (Fall 2018), pp. 7–44, doi.org/10.1162/isec_a_00328. Further references to this article appear parenthetically in the text.
[2] AIC is a popular model comparison metric that balances model likelihood against the number of parameters. Lower AIC values indicate relatively better fits. See Kenneth P. Burnham and David R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed. (New York: Springer, 2002).

International Security, Vol. 44, No. 1 (Summer 2019), pp. 197–200, https://doi.org/10.1162/ISEC_c_00355

three or four.[3] When subjected to a more traditional model selection procedure—an AIC difference of at least three and the simultaneous outperformance of both GDP and CINC scores—the measure yields superior model fits in six of the twenty-four replicated studies. The sample size of twenty-four studies inhibits generalizations, but this reanalysis provides a corrective to the claim that GDP × GDPPC improves fit in the majority of replicated studies.

Second, Beckley's variable introduces potential inferential complications. His proposed measure is equivalent to GDP-squared divided by population. This intuition implies a theory that is quadratic in GDP with the desire to control for population. A more traditional model for such a theory would specify population separately from GDP. The squared nature of GDP further implies that a first order term should be specified. These steps would preserve Beckley's intuition but avoid violations of model hierarchy; furthermore, this specification helps isolate the explanatory work done by population versus GDP.[4]

As evidence, this correspondence replicated each study that utilized a linear capabilities term according to Beckley's approach (i.e., the replacement of the traditional variable with GDP × GDPPC).[5] The models were then respecified with (1) the more traditional battery of population + GDP + GDP-squared, and (2) GDP × GDPPC alongside the variable's square root (i.e., to approximate and control for main effects). Both specifications yield significant fit improvements over the proposed variable in three of the five studies.[6] These improvements are noteworthy, because AIC penalizes models with additional parameters. If one's theory is quadratic in GDP with the desire to control for population, these results suggest that a sounder specification consists of population + GDP + GDP-squared.

Quoting Joseph Nye, Beckley points out that power is like love, "easier to experience than to define or measure" (p. 8). This correspondence echoes Beckley's theoretical critique of GDP and CINC scores. His article's compelling case studies highlight the measure's utility as a standalone indicator of power. To derive more rigorous operationalizations of Beckley's intuition, future work should explore latent factor or principal component manipulations of the same variables—GDP and population. The present inferential results, however, suggest that the proposed variable exhibits a lower goodness-of-fit than Beckley reported, yields an impractical coefficient, and points more naturally to a quadratic specification.

—Caleb Pomeroy
Columbus, Ohio

[3] Beckley mentions as much, but he tallies fit improvements that underperform this threshold. See ibid., pp. 70–71.

[4] GDP-squared is very large relative to population, which introduces quadratic-like effects. This leads to potential hierarchical violations. Breaking population from GDP aids interpretation, because a one-unit increase in GDP-squared/population suggests no straightforward interpretation. On hierarchy, see Peter McCullagh and John A. Nelder, *Generalized Linear Models*, 2nd ed. (London: Chapman & Hall, 1989), p. 89; and Julio L. Peixoto, "Hierarchical Variable Selection in Polynomial Regression Models," *American Statistician*, Vol. 41, No. 4 (November 1987) pp. 311–313, doi.org/10.2307/2684752.

[5] See Beckley, "The Power of Nations," table 3, appendix, doi.org/10.7910/DVN/58KDCM; and this correspondence's replication materials, doi.org/10.7910/DVN/JIJJYE.

[6] AIC improvements include Δ_i = 6, 8, and 148 and Δ_i = 11, 18, and 71, respectively. The other replications display similar fits with the exception of a single case, where GDP × GDPPC exhibits an AIC improvement of 16 over the traditional specification.

Michael Beckley Replies:

I am grateful to Caleb Pomeroy for his helpful response to my article, in which I argue that power should be measured in terms of net stocks of economic and military resources.[1] To support my argument, I show that a net approach more accurately tracks the rise and fall of the great powers and predicts international war and dispute outcomes over the past two centuries than do standard gross approaches. I explain how scholars can apply this net approach in qualitative research and highlight new data from the World Bank and the United Nations that scholars can use in quantitative research.

Unfortunately, these databases only go back to 1990, so they are of limited use for scholars who want to study long-term trends in international relations. To get around this problem, I develop a primitive proxy by multiplying gross domestic product (GDP) by GDP per capita (GDPPC), creating an index (GDP × GDPPC) that gives equal weight to a state's material size and efficiency and for which data are available for many countries going back many decades. I show that this proxy does a better job than standard indicators (GDP and CINC [composite index of national capability]) at predicting the outcomes of great power rivalries and international wars and disputes and improves the in-sample goodness-of-fit in replications of many statistical models published in leading journals over the past five years.

Pomeroy accepts my main conclusion about measuring power and most of the supporting evidence, but raises two issues with the replication analyses. First, whereas I compare Akaike information criterion scores for GDP versus GDP × GDPPC separately from CINC versus GDP × GDPPC (because the replicated studies used either CINC or GDP alone, not both together, in their models), Pomeroy compares GDP × GDPPC against GDP and CINC simultaneously and shows that in only six cases does GDP × GDPPC outperform both by statistically significant margins (i.e., with an AIC score that is at least three points lower). This finding, however, does not seriously challenge my basic claim that GDP × GDPPC is a better proxy for power than CINC or GDP, because GDP outperforms the other two indicators in only three cases and CINC performs best in only two cases.

To be sure, improving the model fit over two competitors simultaneously in six of twenty-four cases is hardly earth-shattering, but it should be remembered that the power variable in the replicated studies is just one of a battery of control variables, so even a substantial improvement in the quality of the power measure may yield a small increase in overall model fit.[2] In other words, it would be unreasonable to expect one change in one variable to radically alter the results of a large number of randomly selected multivariate models in which power was just a control variable. More reasonable would be to expect significant change in particular models in which power is the central variable of interest. In my article, I suggest that the statistical models used in power transition theory and the literature on war and dispute outcomes are ripe for reevaluation. Future research can identify others.

[1] Michael Beckley, "The Power of Nations: Measuring What Matters," *International Security*, Vol. 43, No. 2 (Fall 2018), pp. 7–44, doi.org/10.1162/isec_a_00328.

[2] This point is made explicitly in a forthcoming study that uses the AIC to evaluate measures of national capability. See Robert J. Carroll and Brenton Kenkel, "Prediction, Proxies, and Power," *American Journal of Political Science*, published ahead of print, doi.org/10.1111/ajps.12442.

Second, Pomeroy suggests, for reasons of model hierarchy, that scholars should unpack GDP × GDPC into component parts when using it in regressions. To support his claim, he replicates five studies and shows that using population + GDP + GDP-squared significantly improves model fit over using GDP × GDPPC in three of those studies. Although three out of five cases is rather thin evidence, I have no theoretical issue with his point and appreciate the suggestion.

—Michael Beckley
Medford, Massachusetts

Erratum: On page 77, second line from the bottom of Brooks et al., "The Demographic Transition Theory of War: Why Young Societies Are Conflict Prone and Old Societies Are the Most Peaceful," *International Security*, Vol. 43, No. 3 (Winter 2018/19), "total population" should be "adult population." On page 78, figure 5, the caption "as a proportion of the total population" should be "as a proportion of the population aged 15+."