Verifying the Solution from a Nonlinear Solver: A Case Study: Comment

By RON SHACHAR AND BARRY NALEBUFF*

This paper presents the tale of a replication experiment. The main characters are operating systems, Hessians, scaling, double-peaked likelihoods, and the limits of PC computing. Some of these characters, especially the Hessians, looked scary at first, but turned out to be sheep in wolves' clothing. In other words, the story has a happy ending. To appreciate the twists and turns, we go back and start at the beginning.

Once upon a time, indeed, in the June 2003 issue of the *AER,* B. D. McCullough and H. D. Vinod (2003; "MV" hereafter) set out to test the *AER* replication policy. While many *AER* authors were invited to participate in this replication event, few answered the call. We did. MV singled out our cooperation and honoring of the *AER* replication policy. MV replicated the results in our 1999 *AER* paper (Shachar and Nalebuff, 1999). You might have expected that we would be happy. But we were not.

MV were concerned not only with replication but also with reliability of nonlinear estimation procedures. Specifically, they were concerned that nonlinear solvers can produce inaccurate answers. They believe that this is a systemic problem with empirical research in economics. Thus, they proposed a four-step method to verify the solution from a nonlinear solver. Using data from our paper to illustrate their point they conclude (referring to our 1999 article as "SN"):

> [T]he problem posed by SN has completely exhausted the limits of PC computing, and more powerful computational methods will be needed to analyze this problem properly ... the SN problem, as posed, cannot be reliably solved on a PC.

Our PC took this personally and so did we. We decided to show that the solution is reliable *even* using the procedure proposed by MV. It turns out that our solution passes MV's tests with flying colors. And, MV now share this view as their companion paper confirms. In the process, we also discovered what led them to conclude otherwise in their first paper. Specifically in Section III, we show that a problem that seems beyond the limits of the PC with one scaling of variables becomes solvable with a different scaling. This finding not only demonstrates that our solution is reliable, it also calls for a revision in the approach proposed by MV. Indeed, their new note addresses this issue.

In addition to this statistical insight, we present (in Section IV) an econometric insight that is related to the approach suggested by MV. We show that a problem that has multiple solutions (i.e., a double-peaked likelihood) is recognized to have a single economic solution. The double-peaked likelihood reflects structural restrictions in the economic model and is shown to be equivalent to a single-peaked problem with transformed parameters.

Finally, we present (in Sections V and VI) guidelines that examine the stability and robustness of a solution. These guidelines also indicate whether a local optimum is likely to be also a global maximum.

We begin in Section I with the replication issue. Section II presents the approach suggested by MV and their conclusions about our data. Section VII is our conclusion.

I. Replication? Elementary, Watson

Our 1999 *AER* paper develops and tests a follow-the-leader theory of political participation. There are five tables with estimation results, each providing support for our theory.

* Shachar: Faculty of Management, Tel Aviv University, Tel Aviv, Israel 69978 (e-mail: rroonn@post.tau.ac.il); Nalebuff: School of Management, Yale University, 135 Prospect Street, New Haven, CT 06510 (e-mail: barry.nalebuff@yale.edu). We thank Dmitri Byzalov, David Drukker, Bruce McCullough, James Poterba, and Rick Vinod for helpful comments. We are grateful to Mark Watson who was kind enough to run our GAUSS code on his PC.

MV's study focuses on the *unpublished* results of our first attempt at a structural estimation. These results are presented here in Table 1.[1] Compared to our published results, these estimates offer even stronger support for our theoretical model.[2]

MV were able to replicate these results with TSP. Their TSP estimates (also reported in Table 1) are essentially identical to our GAUSS estimates. However, they were not able to solve this estimation problem using several statistical packages, including GAUSS. (They attribute this lack of success with GAUSS to the operating system they used at the time.) This implicitly raised the issue of whether our results with GAUSS were reliable and replicable.

Thus, we asked several researchers around the globe to try and replicate our Table 1 with GAUSS. Using various computers and operating systems, they have all had the same experience—GAUSS converged to our solution within a few seconds. (It is worth mentioning that our data, code, and starting values are posted on http://www.tau.ac.il/~rroonn/AER.html). Furthermore, the *AER* editor, Professor Ben Bernanke, selected an econometrician of his choice to confirm our results. Professor Mark Watson was kind enough to accept and his results were identical to ours. He wrote to us: "I just ran it ... bing-bang-boom and out popped the answer."

As MV report in their companion paper, with the updated operating system and version of GAUSS they now have the same experience as Professor Watson: "bing-bang-boom." Furthermore, recall that their TSP solution is essentially identical to ours. Thus, it is clear that our result can be replicated.

The first sheep is uncovered.

II. MV Approach and Conclusions

MV present four steps to verify a solution from a nonlinear solver. When they apply these steps to the case study (i.e., our unpublished estimation problem), they find that using their TSP solution: (1) the gradient is zero, and (2) the solution path exhibits the expected rate of convergence. However, they conclude that (3) the Hessian is ill-conditioned, and (4) the Wald test is not appropriate because the quadratic approximation of the likelihood function is not valid.

Based on these findings MV conclude that the reported TSP solution is "at best, a tentative local solution" and that the estimation problem is too large for a personal computer.

MV emphasize that the potential unreliability of the PC solution is a serious issue, especially because "a typical researcher cannot be expected to suspect these pitfalls, let alone verify their existence and know how to handle them properly."

III. Ill-Conditioned Hessian? Not with Proper Scaling

When the condition number of the Hessian is large, the Hessian is ill-conditioned; when the number is small, it is well-conditioned.[3] MV present two thresholds in order to determine whether the condition number is large or small. If the condition number is above 4.5E15 then "the solution is likely to be completely unreliable" and if it is below this value but above 6.7E7 then "the solution should not be accepted uncritically." Based on *their* TSP estimates, MV find that the condition number of the Hessian in the ∞-norm is 1.02×10^{10}. However, based on *our* estimate the condition number in the ∞-norm is 1.4×10^5 (i.e., 1.4E5).[4] In other words, *based on the criteria presented by MV,*

[1] These results are not reported in the *AER* paper. We did not report our initial solution because we found that it depends on two outliers. After moderating the effect of outliers, the new structural estimates gave a weaker but still significant support for our theory. We reported the *weaker* results in Table 9 of our paper. The focus of the MV analysis was not the estimates in Table 9, but rather the results of our initial estimation problem (without the correction for outliers).

[2] Specifically, (1) the parameter ρ is 0.7906 (with standard error of 0.1194), showing that leaders' effort to get out the vote depends on the probability of a tie in the presidential election; (2) the parameter η is 0.1051 (with standard error of 0.0374), indicating that an increase in the voting population decreases leaders' effort; and (3) the parameter S is 0.0672 (with standard error of 0.0294), demonstrating that leaders care about winning their state in addition to winning the general election.

[3] The condition number of a matrix $\mathbf{A}$ is $\|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\|$.

[4] In addition, the Hessian is clearly negative definite.

TABLE 1—THE RESULT OF THE STRUCTURAL ESTIMATION

	Our GAUSS solution Log-likelihood = 1967.07143370			MV's TSP solution
Parameter	Estimate	Standard error	t-value	Estimate
σ_ε	0.1567	0.0048	32.401	0.1567
β_d^o	−1.3367	0.0768	−17.412	−1.337
β_r^0	−0.8897	0.0832	−10.688	−0.8897
ρ	0.7906	0.1194	6.621	0.7906
η	0.1051	0.0374	2.807	0.0000105
S	0.0672	0.0294	2.286	0.0672
$\beta_{\text{Governor's Race}}$	0.6078	0.1551	3.919	0.0608
β_{Rain}	−0.1517	0.0359	−4.222	−0.1517
$\beta_{\text{Jim Crow}}$	−0.3521	0.0351	−10.021	−0.3521
β_{Income}	0.2474	0.0498	4.970	0.2474
β_{Black}	−1.1846	0.0996	−11.891	−1.185
$\beta_{\text{Moved In}}$	−0.0369	0.0075	−4.935	−0.0369
$\beta_{\text{Education}}$	0.3454	0.1025	3.370	0.3454
External states	−0.1829	0.0456	−4.015	−0.1829
1988	0.1713	0.0667	2.567	0.1713
1984	0.2416	0.0640	3.775	0.2416
1980	0.1521	0.0646	2.354	0.1521
1976	0.2667	0.0614	4.340	0.2667
1972	0.2887	0.0546	5.286	0.2887
1968	0.1087	0.0445	2.442	0.1087
1964	0.4661	0.0423	11.009	0.4661
1960	0.3475	0.0357	9.730	0.3475
1956	0.3042	0.0395	7.696	0.3042
1952	0.3627	0.0389	9.320	0.3627
σ_d	0.3765	0.0122	30.755	0.0376
b_0	0.3406	0.0234	14.561	0.3407
Gallup poll	0.5024	0.0281	17.871	0.0050
GNP growth	0.2035	0.0158	12.864	0.0203
Incumbent	0.1325	0.0302	4.387	0.0132
VP candidate's home state	0.1783	0.0846	2.108	0.0178
Presidential candidate's state	0.5649	0.0992	5.693	0.0565
ADA and ACA scores	0.3534	0.0529	6.686	0.0004
Previous vote	0.3249	0.0321	10.130	0.0033
Previous (8 years) vote	0.2487	0.0267	9.318	0.0025
State legislature	0.3642	0.1005	3.626	0.0004
State economic growth	0.6085	0.1678	3.627	0.0061
Deep South (1964)	−0.1446	0.0274	−5.274	−0.1446
Southern Democrat	0.7571	0.0940	8.050	0.0757
South (1964)	−0.8761	0.1673	−5.236	−0.0876
West (1976 and 1980)	−0.7136	0.0957	−7.453	−0.0714
North Central (1972)	0.5544	0.1338	4.143	0.0554
New England (1960 and 1964)	0.6315	0.1181	5.346	0.0631

Note: There are differences in the scales of the variables (between the two solutions) that express themselves through the scales of the parameters.

our Hessian is well-conditioned. Why is our Hessian so much better?

The reason for the huge difference between our Hessian and the one of MV is due to a scaling issue. The following example can shed some light on the effect of scaling on the condition number of the Hessian. Consider the simple linear model: $y_i = \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i$. Assume, for simplicity, that it is known that ε_i comes from a standard normal distribution, and thus the moments of ε_i are not estimated. In this case the Hessian is:

$$\begin{bmatrix} -\sum(x_{1i})^2 & -\sum(x_{1i}x_{2i}) \\ -\sum(x_{1i}x_{2i}) & -\sum(x_{2i})^2 \end{bmatrix}.$$

Let's consider a data set for which the empirical Hessian is $\begin{bmatrix} -0.01 & -0.05 \\ -0.05 & -1 \end{bmatrix}$, and thus, the condition number of the Hessian in the ∞-norm is 147.[5] It is easy to see that this matrix reflects a scaling by a factor of 10 between the measure units of x_1 and x_2. If we rescale x_1 so that it is in the same unit measures as x_2 (i.e., multiply it by 10), then the empirical Hessian would be $\begin{bmatrix} -1 & -0.5 \\ -0.5 & -1 \end{bmatrix}$ and the condition number of the Hessian in the ∞-norm would be 3. In general, the condition number of a Hessian is smaller when the variables are measured in the same units.

In our model the voting population is the one variable that is quite different from the others in terms of its scale. The average voting population is 2,622,000, and the largest is 21,063,000. In the estimation, we measure the voting population in units of 10-millions. Thus, the average number is 0.2622 and the largest is 2.1063—with this scaling the population variable is now quite similar (in terms of scale) to most of the other variables in the data. MV did not employ this scaling approach and measured the voting population in thousands. Thus, the average number is 2,622, and the largest is 21,063. Consequently, the population variable is quite different in terms of its range from the other variables in the data. This, as pointed out above, has significant implications on the condition of the Hessian.

Indeed, when we estimate our model with the scaling of the voting population in thousands (rather than in 10-millions), the condition number of the Hessian is large, and quite similar to the one that MV report.

In any case, our point here is that with proper scaling, even based on the criteria presented by MV, the Hessian of our model is well-conditioned.[6] *Furthermore, since the estimates of the badly scaled problem (executed by MV) are essentially identical to the reliable estimates of the well-scaled problem, we can now conclude that even the solution of the badly scaled problem is reliable.* This means that the approach proposed by MV might lead to a wrong conclusion because it does not take into account proper scaling. In other words, even if the Hessian is not well-conditioned, the solution might still be reliable. The statistician's job is not done. She should search for the optimal scaling and only then the condition number of the Hessian can be used to verify the solution. Indeed, in their companion paper MV update their approach to address this issue. Furthermore, in the conclusion we extend our discussion of the scaling issue.

The second sheep is uncovered.

IV. Is the Likelihood Bimodal? Yes Does It Matter? No

The Wald statistic is often used to test a hypothesis. However, its accuracy and reliability depend on the quadratic approximation of the likelihood function. If this approximation is not reasonable, inference should rely on the likelihood values (for example, likelihood ratio tests, as we did in our original study for the parameter S). The fourth step presented by MV focuses on this issue. They suggest profiling the likelihood in order to assess the adequacy of the quadratic approximation.

Figure 1(b) in their study presents the profile of the likelihood with respect to the parameter η (the parameter that multiplies the voting population variable). This figure, presented below as Figure 1, is a result of the following exercise: set η at a specific value and estimate all the other parameters of the model, then set η at a different value and reestimate, etc.

[5] The ∞-norm of a matrix is $\max_{1 \le r \le R}(\sum_{c=1}^{C}|a_{r,c}|)$ where R is the number of rows in the matrix, C is the number of columns, and $a_{r,c}$ is the element in the r row and c column (i.e., the maximum absolute row sum). Thus, the ∞-norm of the Hessian above is 1.05. The inverse of this Hessian is $\begin{bmatrix} -133\frac{1}{3} & 6\frac{2}{3} \\ 6\frac{2}{3} & -1\frac{1}{3} \end{bmatrix}$. Thus, the ∞-norm of the inverse is 140. The condition number, as mentioned above, of a matrix $\mathbf{A}$ is $\|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\|$. Thus, the condition number of the Hessian is $1.05 \cdot 140 = 147$.

[6] David M. Drukker and Vince Wiggins (2004) reach the same conclusion and suggest further research on when the condition number of the Hessian is too high.

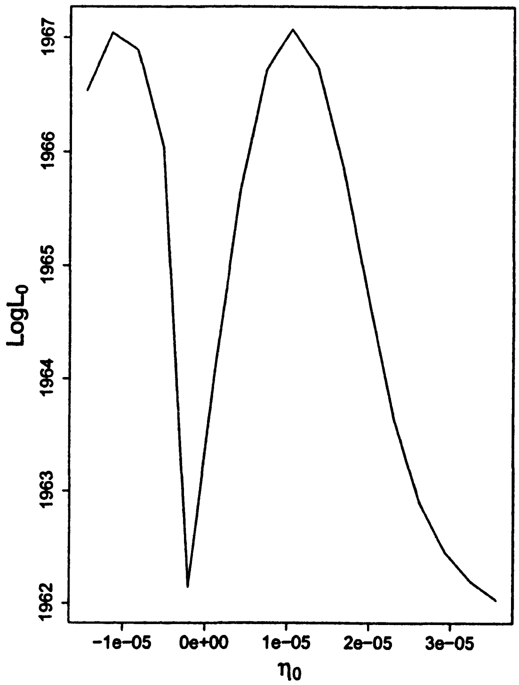

FIGURE 1. FIGURE 1(b) IN THE MV STUDY

At first glance, the likelihood does not seem quadratic. Based on this evidence, MV conclude that the quadratic approximation is not valid. It turns out that Figure 1 is too coarse and thus cannot reveal the interesting effect that η has on the likelihood.

MV's figure is based on eight values of η. In Figure 2 below we do the same exercise using 420 values (and a wider range) of η.

There are two immediate conclusions from Figure 2 about the likelihood function: (1) it has two peaks, and (2) it is completely symmetric around zero.

Should we be disturbed by the two peaks? No. Our structural model implies that the likelihood function is *symmetric* around zero and should have two peaks. Thus, Figure 2 is exactly what one should expect. Specifically, the parameter η appears only once in the likelihood function, and in this single appearance it is multiplied by the parameter ρ. Furthermore, when ρ does not multiply η, it always appears as ρ^2. This can be seen from the unnumbered equation on page 541 of our 1999 study:

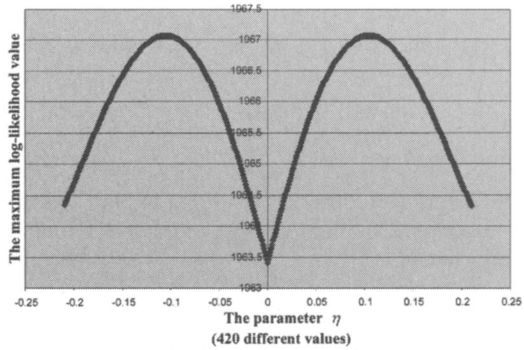

FIGURE 2. PROFILE OF THE LIKELIHOOD WITH RESPECT TO η

$$\text{(1)} \quad \rho E_{jt}^{*} = \rho^{2}(S + R_{jt})\alpha_{1}(1 - \alpha_{1}) \times \phi\left(\frac{\alpha_{1} - \bar{d}(\zeta_{jt})}{\sigma_{d}}\right) - \rho \cdot \eta \cdot N_{jt}.$$

(Note that ρ and E_{jt}^{*} appear only once in the likelihood function as ρE_{jt}^{*}.) Thus, the theory predicts that $\rho = 0.7906$ and $\eta = 0.1051$ should lead to exactly the same likelihood as $\rho = -0.7906$ and $\eta = -0.1051$. Furthermore, these two pairs of parameters should have exactly the same behavioral effects (i.e., the effect of predicted closeness, electoral votes, and voting population on the political participation are exactly the same).

Economically, it makes no difference if ρ *and* η *are both positive or both negative.* What matters is only if they are the same sign or not and if they are different from zero.

Thus the two peaks represent one economic solution. And the adequacy of the quadratic approximation should be examined either for the positive values of η or for the negative values.

The third sheep is uncovered.

Figure 2 also demonstrates that the likelihood is not exactly quadratic. The highest nonnegative value of η is 0.210 and the lowest is obviously zero. Since the estimate of η is approximately 0.105 (exactly in the middle between 0.210 and 0), the likelihood is not exactly quadratic if the values of the function at 0.210 and 0 are not the same. Figure 2 shows that the values differ, and thus the likelihood is not exactly quadratic.

However, the likelihood does not have to be

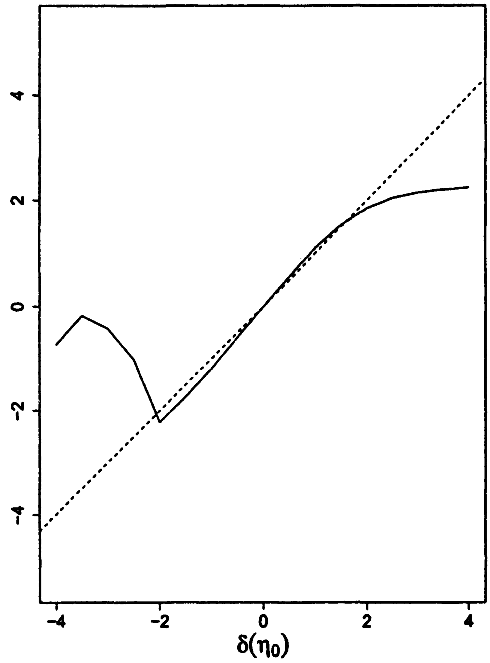

FIGURE 3. FIGURE 1(a) IN THE MV STUDY

exactly quadratic in order to use the Wald test. It should be approximately quadratic locally (near the estimates). And it is. As MV point out, it is easier to identify deviations from linearity than from quadratic behavior. Thus, we also replicate their Figure 1(a), that is based on the following transformation. The value on the horizontal axis is the deviation of the η from the estimate $\hat{\eta}$ in units of standard deviations $\left(\text{i.e.,}\frac{\eta-\hat{\eta}}{\sqrt{\text{Var}(\hat{\eta})}}\right)$, and the value of the vertical axis is $\text{sign}(\eta - \hat{\eta})\sqrt{2[\text{LogL}(\hat{\eta}) - \text{LogL}(\eta)}$. If the resulting graph is a linear line with slope one, the likelihood is exactly quadratic.

Figure 3 presents the relevant MV plot (Figure 1(a) in their original study), and Figure 4 the plot based on our analysis. The kink on the left-hand side in Figure 3 is due to the double-peaked likelihood, and thus should be ignored.

As expected the graph is not linear for extreme values of η. However, as pointed out above, for the quadratic *approximation* to hold, it is only required that the graph should be almost linear. Furthermore, we are interested in the behavior of the curve near the estimate. It

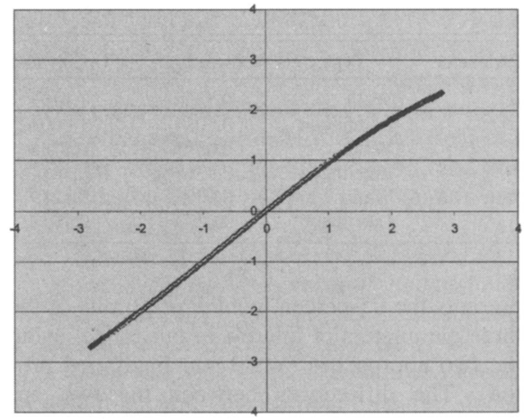

FIGURE 4. LINEARIZATION OF THE QUADRATIC APPROXIMATION WITH RESPECT TO η

turns out that locally, the quadratic approximation is adequate. Specifically, within an interval of one standard deviation it is linear with a slope of one, and within an interval of two standard deviations it is still linear with a slope of one for values of η that are lower than $\hat{\eta}$. Thus, although the likelihood is not exactly quadratic, it is approximately quadratic. This means that inference based on the Wald statistic is quite reliable. Indeed, the Wald test and the LR test yield a similar conclusion with respect to the hypothesis that $\eta = 0$: the relevant χ^2 statistic is 7.34 suggesting that the hypothesis can be rejected even at the 0.67-percent level; and the t-statistic is 2.807 suggesting that $\eta = 0$ can be rejected even at the 0.25-percent level.[7] This means that we would have concluded that the effect of the voting population on leaders' effort is positive at the 1-percent significance level with both the Wald test and the LR test.

Furthermore, the confidence intervals based on the two approaches are quite similar. Specifically, for a significance level of 95 percent, the Wald confidence limits are 0.0315 and 0.1786, and the profile likelihood confidence limits are 0.0313 and 0.1874. The lower limits are essentially the same, and the upper limits differ a bit. This small difference is obviously related to the decrease in the slope of the line in Figure 4 beyond the first standard deviation. Table 2

[7] We are using a one-sided test, because, as discussed above, the sign of the parameter is not important.

TABLE 2—95-PERCENT CONFIDENCE LIMITS

Parameter	Type	Lower limit	Upper limit
ρ	Wald	0.5560	1.0253
	Likelihood	0.5084	1.0008
η	Wald	0.0315	0.1786
	Likelihood	0.0313	0.1874
S	Wald	0.0094	0.1249
	Likelihood	0.0273	0.1932

presents the 95-percent confidence limits of the three parameters of interest in our model using the two approaches (Wald and likelihood profile). The differences between the two approaches are almost meaningless from an economic point of view. More importantly, all three parameters are different from zero at the 1-percent level according to both approaches.[8]

Thus, the quadratic approximation is kosher, and all the parameters of interest are strongly different from zero.

V. The Lesson from Reparameterization

The four steps suggested by MV are designed to verify that a solution from a nonlinear solver is indeed a maximum. As illustrated above, our solution passes this test with flying colors. However, the solution might still be a local maximum, not a global one. In the following two sections we present several practical ways to examine the stability and robustness of a solution. A solution that passes these tests is likely to be a global maximum.

In this section, we take advantage of the structural restrictions imposed by the theory [as illustrated in equation (1)]. Note that equation (1) above can be rewritten as:

$$(2) \qquad (\beta_1 + \beta_2 \cdot R_{jt})\alpha_1(1 - \alpha_1) \times \phi\left(\frac{\alpha_1 - \bar{d}(\zeta_{jt})}{\sigma_d}\right) + \beta_3 \cdot N_{jt}$$

where the three structural parameters of interest (ρ, η, and S) are replaced with three reduced-form parameters (β_1, β_2, and β_3) and $\beta_1 = \rho^2 S$, $\beta_2 = \rho^2$, and $\beta_3 = -\rho\eta$. Obviously, it is easy to solve the structural parameters from the βs. It is clear that the ML estimates of the model with equation (2) must give the same log-likelihood as the ML estimates of our model with equation (1). Table 3 presents the estimates of these three βs and the log-likelihood.[9] As expected, the log-likelihood of the reparameterized problem is identical to the log-likelihood with the structural parameters. Furthermore, the solution of the structural parameters from these βs is identical to those reported in Table 1.

TABLE 3—THE RESULT OF THE REPARAMETERIZED ESTIMATION
Log-likelihood = 1967.07143370

Parameter	Estimate	Standard error	t-value
β_1	0.0420009111778	0.0106	3.972
β_2	0.625121221821	0.1888	3.311
β_3	−0.0830737197820	0.0306	−2.718

Note: $\hat{\rho} = \sqrt{\hat{\beta}_2} = 0.7906$; $\hat{\eta} = \frac{\hat{\beta}_3}{-\hat{\rho}} = 0.10507$; and $\hat{S} = \frac{\hat{\beta}_1}{\hat{\rho}^2} = 0.067188$.

This exercise demonstrates that the estimates in Table 1 are robust and stable, and shows that these estimates maximize the likelihood even when the estimation problem is reparameterized.

Furthermore, Figure 5 presents the log-likelihood as a function of β_3, the parameter that multiplies the voting population, and Figure 6 presents the linearization of the quadratic approximation. The quadratic approximation is beautiful. Thus, although the likelihood is only approximately quadratic with respect to η, it is exactly quadratic with respect to β_3. This means that reparameterization is a useful approach not only in order to examine the robustness of a solution, but also in improving the quadratic approximation.

Another way to examine both the robustness of the model and the reliability of its inference is by comparing the values of the likelihood in the reparameterized model when β_3 is set at zero and in the structural model when η is set at zero. These two values are supposed to be the same. And they are (the log-likelihood in both cases is exactly 1963.401549). Thus, the χ^2

[8] According to the Wald profile the S parameter is different from zero only at the 2.5-percent level.

[9] The other parameters in this estimation are identical to those reported in Table 1.

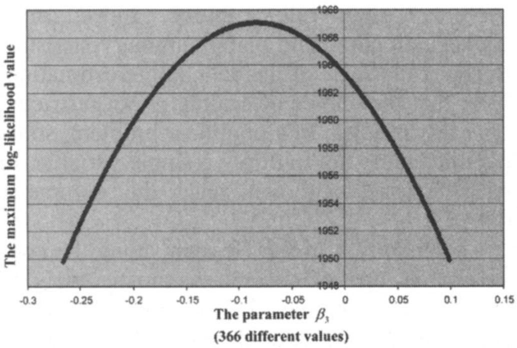

FIGURE 5. PROFILE OF THE LIKELIHOOD WITH RESPECT TO β_3 (REPARAMETERIZED MODEL)

statistic of the hypotheses $\beta_3 = 0$ (in the reparameterized model) and $\eta = 0$ (in the structural model) are identical.

VI. Indications that a Solution is a Global Maximum

Reparameterizing the estimation problem is one of the ways that applied researchers can use in order to examine the stability and robustness of their estimates. Here are a few additional methods that we have used before publishing our 1999 article:

- We started the estimation procedure from a wide range of starting points. The initial starting points were based on the ordinary least-squares (OLS) estimations that serve as nonstructural tests. Furthermore, the estimation procedure converged to the exact same solution even when it started from random starting values.
- We experimented with the tolerance level, and other options of the GAUSS procedure.
- We used Monte Carlo experiments in order to test our estimation procedure and to analyze the empirical model. We created random data sets using our structural model and then verified that the estimates are consistent with respect to the parameters used to create the data.
- We compared the structural estimates to those found via OLS nonstructural estimation and found similar economic effects.
- The success of the out-of-sample predictions (as in Table 11 of the 1999 article) can also serve as test of the empirical model.

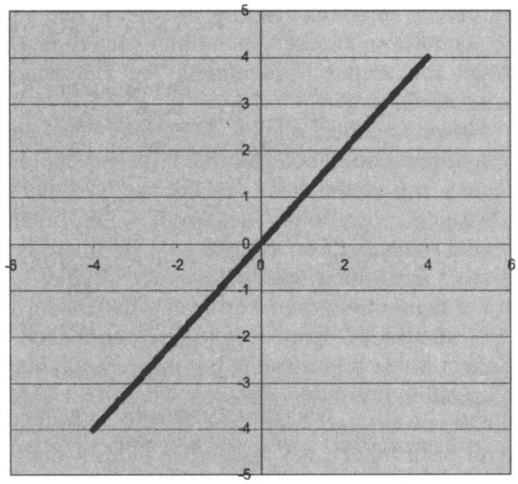

FIGURE 6. LINEARIZATION OF THE QUADRATIC APPROXIMATION (β_3)

The evidence that a solution is stable and robust also indicates that it is likely to be a global maximum.

VII. Conclusion

MV raise two important issues in their paper: (1) the need to verify the solution of a nonlinear estimation procedure, and (2) the need to reform the *AER* replication policy. We, too, are concerned with both issues.

We start with the *AER* replication policy. Recently the *AER* changed its policy and now requires authors of accepted empirical papers to provide all programs and data files for posting on the *AER* Web site as a precondition of publication. We suggest that the journal go one step further and assign a special editor to handle replication issues. If authors know that there is a special editor overseeing the process, their tendency to *fully* cooperate would increase. A further way to motivate authors is to provide them an acknowledgement even if their results are replicable.

Turning to verifying solutions from nonlinear estimations, a central element in MV's method is the condition of the Hessian. The rationale behind MV's suggestion is simple: it is well known that when a model is not identified, the Hessian is singular. Most statistical packages inform the user about such a problem. In such cases, the Hessian is clearly ill-conditioned and its condition number is infinity. In contrast, when the model is identified, the Hessian is nonsingular.

However, the simple dichotomy between "identified" and "not identified" is misleading. Depending on the data, a model can be "poorly identified" or "well identified." The linear model sheds light on this point. When there is a perfect linear relationship between variables (or some combinations of variables), the model is "not identified." Even when there is not such a perfect linear relationship, but rather just a high correlation (i.e., two or more variables, or combinations of variables, are highly correlated with each other), the researcher faces a multicollinearity problem. In such a case, the estimates are not accurate, and the model is "poorly identified." One signal that such a problem exists is that the condition number of the Hessian is large, though not infinity.[10]

This leads to the question: when is the condition large enough so that the researcher should be worried about the accuracy of its estimates. MV provide us with thresholds to address this question. We have shown that the condition number of the Hessian is sensitive to the scaling of the variables. Specifically, a Hessian that is ill-conditioned with one scaling is well-conditioned with a different scaling (without changing the solution). This means that the method suggested by MV applies only to well-scaled problems.

What are the proper scales? In the text we demonstrated that the condition number of a Hessian can be greatly reduced when the variables are measured in the same units. This can serve as a good "rule of thumb." Indeed, most packages recommend using summary statistics to check the scale of the data before estimation. Note that the choice of scaling is not restricted only to variables. In a nonlinear problem, some parameters do not multiply a single variable. In such a case, one can scale the parameter directly.[11]

The condition of the Hessian should be considered as one of several diagnostic tools of an estimated model. If the Hessian is ill-conditioned, one should study the variance-covariance and the correlation matrixes of the parameters to locate the source of the problem. Graphing the likelihood as a function of two parameters can be a useful tool in identifying the source of the problem.

Finally, we are pleased that MV confirm that our original problem was properly scaled and that the estimation results are both replicable and reliable.

[10] An alternative term for "poorly identified" is "weak identification." The latter has been widely used with respect to the IV estimator (see, for example, James H. Stock and Motohiro Yogo, 2002). At least in one aspect the multicollinearity in the linear model, the problem that MV focus on and the IV weak identification issue are similar. In all these cases, identification depends on a full rank condition. In the simple linear model, $X^T X$ must be of full rank. The IV estimator requires that $W^T X$ is of full rank, where W is the matrix of instruments. In the nonlinear model, the Hessian must be of full rank. We thank David Drukker for pointing out this analogy.

[11] For example, instead of estimating β, one can estimate γ, where $\beta = s\gamma$ and s is the scale. This means that anywhere that β appears, it is replaced by $s\gamma$ (e.g., $0.1 \cdot \gamma$ if the chosen scale is 0.1). When the estimation is done, in order to recover the parameter of interest the researcher should calculate $s\hat{\gamma}$ and the standard error, accordingly.

REFERENCES

Drukker, David M. and Wiggins, Vince. "Verifying the Solution from a Nonlinear Solver: A Case Study: Comment." *American Economic Review,* March 2004, *94*(1), pp. 397–99.

McCullough, B. D. and Vinod, H. D. "Verifying the Solution from a Nonlinear Solver: A Case Study." *American Economic Review,* June 2003, *93*(3), pp. 873–92.

Shachar, Ron and Nalebuff, Barry. "Follow the Leader: Theory and Evidence on Political Participation." *American Economic Review,* June 1999, *89*(3), pp. 525–47.

Stock, James H. and Yogo, Motohiro. "Testing for Weak Instruments in Linear IV Regression." National Bureau of Economic Research (Cambridge, MA) Working Paper No. t0284.