

**Note: Alternative Treatments of Congestion in DEA -- A Response to the Cherchye, Kuosmanen
and Post Critique**

By

**W.W. Cooper* (Corresponding Author)
Bisheng Gu*
Shanling Li+**

***University of Texas at Austin
*Graduate School of Business
Austin, TX 78712-1174, USA**

**+McGill University
Faculty of Management
McGill University
Montreal, Quebec
Canada H3A 1G5**

May 2000

Note: Acknowledgement is due the IC2 Institute of the University of Texas at Austin for support of the research of W.W. Cooper. An NSERC grant of Canada helped to support the research of Shanling Li.

1. Introduction:

We congratulate EJOR on its constructive resolution of the impasse that developed between the referee and the authors of the paper referenced in the above title. The original version of the paper was accepted without qualification by one referee and rejected by the other referee (apparently 3 in number) also without qualification. Subsequent exchanges and revisions of the paper revealed a difference in the philosophy of this referee and the authors (also 3 in number) whereupon the editor, Jyrki Wallenius, invited the authors and the referee to participate in the open exchange to which we now turn.

Before undertaking our response, we want to make clear that we are not recommending abandonment of the FGL (Färe, Grosskopf, Lovell) approach to congestion, as is asserted by CKP (Cherchye, Kuosmanen and Post). On the contrary we are recommending that its deficiencies should be repaired so that its use can be improved upon and widened. As noted in the concluding paragraph of our article, "Much has been accomplished by this [FGL] approach.. and we believe that much more can be accomplished with repair of its deficiencies." Indeed, we have elsewhere offered suggestions for repairing the deficiencies we have noted. Furthermore -- while awaiting completion of these repairs -- we have also suggested procedures for determining whether such deficiencies are present and how precautions might be taken in any particular application of the FGL approach. See our concluding comments, below, in this note.

We can best make clear the nature of our "philosophical" (=philosophy of science) differences with CKP by citing a long tradition in economics which is reflected in some of the DEA (and DEA like) literature. Starting with Afriat (1972), developed in detail by Hanoch and Rothschild (1972), continuing through Diewert and Parkan (1983), and on into the textbook

literature, as in Varian (1993), the efforts in this part of the literature are directed to determining whether bodies of data are suitable for use in economics. When this is not the case the data (including important bodies of government data) are to be rejected. The comments by CKP are in this tradition. Our position, on the other hand, is that it is the job of theory to explain the data. When the two do not conform, it is the theory that should be altered. In part, as already noted, this is what we are suggesting for FGL.

Remark: This does not prevent adjusting the original data (e.g., as in index number adjustments) provided the nature of the adjustments is explicitly noted and the reasons for making the adjustments are set forth clearly.

2. Differences in Philosophy:

We now proceed to document this difference in philosophy as follows. First we note that the major criticisms offered by CKP generally rest on their statements that we do not conform to assumptions posited in economics as well as in FGL or embedded in their models. Second, they fault us because we fail to conform to the original motives of FGL in developing a measure of "structural efficiency" with, presumably, decomposition into congestion and other types of inefficiency representing only a by-product. Third, and finally, CKP argue that our concepts are ill-defined because they are not founded on quantitatively formulated definitions.

Starting with the third criticism we provide responses in reverse order as follows. Leaving aside the peregrinations followed by FGL in developing their definitions, as tracked by CKP (immediately after their reference to Definition 1), we focus on the fact that we do supply all of the quantitative measures that we believe are needed to deal with (input) congestion and its evaluation in DEA. Reference to Appendix A in our article shows that we have provided an overall measure which represents (i) the average proportion of congestion in all of the input

amounts as well as (ii) the average proportion of technical inefficiency for each DMU.

Reference to our formula (10) shows that for each input, for every DMU, we also decompose the total inefficiencies represented by each of the input slacks into a congesting amount and a technical amount. Finally the output losses due to congestion for each DMU are then easily derived by noting the differences in their values in the rims of our first-stage and second-stage models.

The statement that FGL are interested [only] in particularly severe forms of congestion is [only] inferred by CKP. In fact FGL supply an overall measure which fixes only upper or lower bounds on the amount of congesting input or congested output, whatever its amounts and whether it is severe or not. See the Appendix to Brockett *et al.* (1998). The algorithms supplied by FGL involve all possible partitions of the data if one wants to go further into estimating the amounts of congestion in each input or output. In any event what is meant by "particularly severe congestion" is left unspecified by CKP. See the definitions we supply in the opening discussion of our article where we note that "congestion" is a "particularly severe form of technical inefficiency" where, as we also note, the two can be distinguished by the property that the former, but not the latter, requires output to fall with input increases.

We can further clarify the criticism offered by CKP in their section 3 by inferring (from their referee reports) that they are saying that the correct way to treat the topic of our paper is by analytically formulated axioms and set theoretic characterizations. Our preference, however, is for verbal definitions (like ones we use) in order to facilitate contact with potential users. We also prefer algebraic characterizations which provide immediately implementable formulations.

In any case, we believe that our definitions, together with our subsequent developments, provide what is needed both for immediate use and further research.

Turning next to the second of the two major criticisms offered by CKP --viz., that we fail to conform to the "original" motives of FGL --we can respond by noting that (valid or not) reference to the "original motives" in FGL, whatever they might be, does not address the issue of whether results obtained from their congestion formulas are correct. Indeed, as shown in our paper, this is not the case when applied to bodies of data like we discuss in our article. The possibility of incorrect inferences is important because the FGL approach to congestion has been (and continues to be) extensively used in empirical work. This includes uses by FGL who, presumably, are aware of their motives for developing these approaches. In fact, what CKP refer to as "our example" overlooks our statement (in the article) that this example was taken from FGL (1985) and in addition -- as noted in our discussion of Tables 1 and 2 -- this example was used by FGL (1985) to illustrate their approach to congestion. The fact that congestion is found to be present when (as we show) the data do not support this conclusion seems to be of little or no interest to CKP who defend the use of FGL, despite this result, either by (a) introducing assumptions to explain this result or by (b) using properties of the model rather than the data to justify waiving aside this result.

Turning to the third charge levied against us by CKP -- i.e., we do not conform to the assumptions of economics (and FGL) -- we plead nolo contendere in the common law sense. See Black (1979). We are, indeed, willing to alter economic (or other) theories when they do not fit the facts. However, we are not entirely comfortable with this response because it is not always clear what CKP mean with their use of terms like "stylized (=theoretical?) facts," as in

their discussion of the "law of variable proportions" in section 2 of their critique. Their example of traffic congestion as measured by "marginal transportation output" is also stated in unusual terms. The usual statement, in traffic engineering, is in terms of "transit times" like those noted in the discussion of the Chicago Area Transportation (CATS) study which is given on pp. 785 to 793 in Charnes and Cooper (1961). This study, we might note, was mounted under joint Federal, State and Local Government sponsorship (shortly after World War II) and charged with responsibility for formulating plans for a street network for the City of Chicago. A. Charnes and one of the authors of this paper (Cooper) were retained as consultants to help CATS develop alternative approaches when the "one-road-at-a-time" analysis with its accompanying assumptions usually used for transportation analysis (at that time) gave incorrect results.

Remark: It may be of interest to note the uses of Wardrop's principle in a manner akin to Pareto-Koopman optimality (as that term is used in DEA) provided a key for guiding the simulations and network optimizations used in these studies. See the above cited reference to Charnes and Cooper (1961).

Trouble arises with terminology like "marginal transportation output" as used by CKP, because a street widening like the one they describe is likely to lead to a change in the mix of trucks and autos as drivers seek to minimize their origin-to-destination travel time -- e.g., as stated in Wardrop's principle for use in traffic planning (like the one undertaken by CATS). Hence the term marginal transportation output is ambiguous -- and especially so if interpreted as exhibiting the returns to scale that CKP assert since mix changes are not allowed in economic usage of the returns-to-scale concept. In any case we prefer the example of too many miners in an underground mine -- where output can be increased by reducing the number of miners. If a more "modern" example is wanted, one can think of the use of "flaming" to congest an address

on the internet -- which increases the message input to this address in order to reduce its output of desired (or desirable) messages.

The "idling assumption" referred to by CKP in section 4 of their critique is clearly not applicable to their example of traffic flow or to our example of "flaming." It is appropriate for our example of miners in an underground mine. Unfortunately, however, CKP fail to provide any suggestion of how quantitative estimates of idling are to be obtained when the needed information is not reported in the data (the usual case) or even how the validity of this assumption is to be determined.

3. Counter Examples:

We now turn to the 2 supposed counter examples erected by CKP. In both cases the counter examples rest on assumed behavior and not on any data. Thus, in example 1 it is assumed that output will fall from 2 to 0 for DMU₇ if its input is increased from $x_1 = 4$ to $x_1 = 4 + \varepsilon$, $\varepsilon > 0$. Thus, by assumption, CKP turn the data for DMU₇ into a boundary point for the set of production possibilities. Activity beyond this point is assumed to lie outside the set of production possibilities. Hence it is assumed that output will fall from $y=2$ to $y=0$ when input is employed beyond this boundary. Our response is that the counter example is, at best, a reflection of their assumptions. As evidence we note that none of what CKP say appears anywhere in the discussion of this example by FGL who are its authors.

The second counter example provided by CKP is based on the assumption that DMU lies on the congested part of the frontier as drawn in our Figure 2. However, the same result would be secured if G were moved into the interior along the ray to the origin while staying inside the congested region of this Figure. CKP then claim that our results are spurious because they do

not provide results that conform to their assumption! One may therefore wonder why FGL, as well as we, turned to the data in these examples since we could (like CKP) equally well simply discuss the properties of our models without reference to numerical data.

The discussion of the data by FGL and the results we secured in example 1 are brushed aside by CKP along with our discussion. This, we believe, displays the attitude toward data that we identified with CKP at the start of this note. In an earlier version of our paper we attempted to distinguish our approach as "data based," in contrast to FGL which we characterized as "theory based." We believe this is justified by statements like the following in FGL (1985, p. 187) where they summarize the motives for book as follows,

"The basic purpose of the research [reported in this book] has been to develop a coherent theory of production in the presence of inefficiency [and thereby extend the economic theory of production]."

We then went on to say that results other than those exhibited by the data could be admissible provided the nature and the reasons for differences from the evidence were made explicit. To meet the referee's objection to this way of distinguishing between these two approaches, this material was all removed from this earlier version of our paper because we did not think it was important for our development. It is nevertheless pertinent here because no such explicit statements are provided by FGL in the discussion of their example.

4. Numerical Examples:

a. Counter Example 1:

CKP do not provide any results that might serve to check whether their assumption that uses of our models (7) and (8) will fail to show congestion when the data are manipulated in the manner they prescribe. To check the validity of this assumption we therefore replace $x_1 = 4$ for

DMU7 with $x_1 + \varepsilon = 4.25$ and $y=2$ with $y=0$. The latter assumption produces ϕy with $\phi \rightarrow \infty$ in our model (7). The expression $\infty \cdot 0$ is not well defined -- either in mathematics or economics -- so we impose a bound $\phi \leq K$ where K is an arbitrarily large real number. This then gives $\phi \cdot 0 = 0$ -- which is consistent with the CKP assumption that only output zero is possible beyond this boundary.

Remark: Alternatively we could argue that $\lim_{\phi \rightarrow \infty} \phi y = 0$ since $y=0$ occurs as soon as the boundary is crossed. Hence $y=0$ occurs before ϕ reaches ∞ .

Using the thus altered data, maximization of the slacks, as required for an optimum in our model (7), yields

$$\lambda_3^* = 1, s^{+*} = 2, s_1^{-*} = 2.25, s_2^{-*} = 0.25$$

and all other $\lambda_j^* = 0$. Application of our formula (9) produces the new values

$$\lambda_6^* = 1, \delta_1^{-*} = 1, \delta_2^{-*} = 0.25$$

and all other variables zero. Formula (10) in our paper thus gives

$$\begin{aligned} s_1^{-C} &= 2.25 - 1 = 1.25 \\ s_2^{-C} &= 0.25 - 0.25 = 0. \end{aligned}$$

Hence, contrary to the assumption of CKP, our approach does not fail to identify the congestion that occurs (on their assumption). Moreover, we identify x_1 as the congesting input, and $s_1^{-C} = 1.25$ as the congesting amount of this input which, together with technical inefficiency in the amount $\delta_1^{-*} = 1$, accounts for all of the inefficiency. Notice, in particular, that $s_1^{-C} = 1.25$ is associated with the fall in output from $y=2$ to $y=0$ while $\delta_1^{-*} = 1$ represents the remaining

inefficiency because it represents an input amount in excess of what is needed to produce this value of $y=2$, but is not associated with the reduction in input that is accounted for by s_1^{-c} .

Turning in the opposite direction we can observe that replacing $x_1 = 4.25$, by $x_1 = 4$ retrieves the original data set where $y=2$ is to be restored. As before, congestion is then found not to be present in the performance of DMU₇ by our model. Furthermore (and finally) all values between 4 and 4.25 will evidence congestion when our model is used. We therefore conclude that the assumed counter example of CHP fails to hold in all respects.

Remark: A use of the unified additive model that was introduced in Cooper, Seiford and Zhu (2000a) would have removed any trouble in treating $y=0$. In fact, the first and second stage approach we used in the above discussion conforms to a use of this unified additive model. Hence a use of this alternative model produces the same result and therefore confirms the presence of congestion. (This helps to show the advantage of having alternative models available since, inter alia, they can be used to check whether the same, or similar, results are secured. See the discussion of the need to check for "methodological bias" in Charnes, Cooper and Sueyoshi (1988a, b))

b. Example 2:

We have already noted that DMU G need not lie on the "congested part of the frontier" as assumed by CKP because the same result would hold if G were moved into the interior of this region. We might also note that the difference in returns to scale assumptions which CKP emphasize in their critique is not pertinent because (a) the solution to our model (11) satisfies the convexity requirements for a variable returns to scale model and (b) the theorem we cite is applicable in either case. See Cooper, Seiford and Zhu (2000b).

The reasoning of CKP in this second example is also faulty since it does not conform to the requirements in economics -- where returns to scale is unambiguously identified only with behavior on the frontier. The movements on the ray from G in Figure 2, however, occur below

the surface of the corresponding 3-dimensional surface. This includes movement in the congested (hence inefficient) region where, even on the surface, no tradeoff between inputs and outputs is required. Both may be improved simultaneously without any tradeoff so the economic concept of returns to scale is not applicable.

We can draw further implications for the analysis presented by CKP by returning to the discussion of Figure 2 in our article. As noted in our article, the ray from the origin to G passes directly under R, the point of maximum output. The input reductions which are identified by our model as being needed to reach R are found to yield a 10-fold increase in output. The point identified as satisfactory by CKP, however, shows no output reduction at all! Hence output augmentations accompanied by input reductions in all parts of this set of production possibilities are completely concealed from view. This seems empirically unsatisfactory, at least to us, no matter how rigorously established (or required) by an underlying theory.

Returning to the issue of philosophical differences, CKP appear to be satisfied by this outcome because it conforms to the properties that they assign to the FGL models. For our part, we prefer what is done in standard medical practice: When one diagnostic technique, an x-ray, say, does not identify the sources of trouble, recourse is had to other techniques (and models of analysis) such as CAT scans and MRI imaging.

5. Conclusion:

This leaves open the question of "where do we go from here?". CKP issue a "call for a more systematic treatment of congestion." Referring again to one of their referee reports, we take this to mean that they want a complete redo of the subject up to (and including) full-scale statistical formulations. This is likely to take a long time and lead to a monograph rather than an

article (or even several articles) in order to report its results. However, in the meantime, we would be willing to accept whatever progress might be made -- piecemeal or not. Such progress could take the form of clarifications like those exhibited in the exchanges with Färe and Grosskopf that we cited in the Reference. It could also include positive responses to the thus clarified issues. An example is provided in Färe and Grosskopf (2000). In response to the showing by McDonald (1996) (with "one or two examples!") that the FGL decomposition of technical inefficiency into scale and congestion components is not commutative, Färe and Grosskopf (2000) provide guidelines that can help in selecting the order in which these decompositions are made. This constitutes progress, we think, at least for those of us who are concerned with obtaining valid empirical applications of these ideas.

Analysis of model properties, as in CKP, is necessary both for valid inferences and knowledge of model limitations, but it is not sufficient for purposes of empirical use. A great deal of the progress in DEA has come about from its uses in empirical applications and it is our belief (and hope) that this will continue since it is also our belief that model development and applications should be a joint endeavor. This can provide opportunities for the two to interact in ways that lead to deeper understanding of relations between different models as well as deeper understanding of the limitations (and potentials) of DEA. The discussion in this exchange, as well as in our article, are examples of the latter. An example of the former is the equivalences established between FGL and our approach which are presented in Cooper, Seiford and Zhu (2000a and b). This interaction between empirical use and theoretical development can lead to new types of models such as the unified additive model introduced in Cooper, Seiford and Zhu (2000a). It can also lead to insights and suggestions for improving the uses of existing models.

See, for instance, the suggestion in Cooper, Seiford and Zhu (2000b) that non-zero output slack be attended to in the solutions obtained by FGL as a readily available way of identifying whether output reductions occurred -- thereby avoiding erroneous inferences like those effected by FGL in example 1 where such non-zero output slacks cannot occur. Similarly, examination of whether non-zero input slacks occur can help to identify whether the law of variable proportions is operative. See the discussion in Cooper, Seiford and Zhu (2000b).

REFERENCES

- Afriat, S.N. (1972) "Estimation of Production Functions" *International Economic Review* 13, 568-598.
- Black, H.C. (1979) *Black's Law Dictionary, 5th Edition. Definitions of the Terms and Phrases of American and English Jurisprudence, Ancient and Modern* (St. Paul, Minn. West Publishing Co.)
- Brockett, P.L., W.W. Cooper, H.C. Shin and Y. Wang (1998) "Inefficiency and Congestion in Chinese Production Before and After the 1978 Economic Reforms," *Socio-Economic Planning Sciences* 32, pp. 1-20.
- Charnes, A. and W.W. Cooper (1961) *Management Models and Industrial Applications of Linear Programming* (New York: John Wiley & Sons, Inc.).
- Charnes, A., W.W. Cooper and T. Sueyoshi (1988a) "Goal Programming/Constrained Regression Review of the Bell System Breakup," *Management Science* 34, pp. 1-26
- Charnes, A., W.W. Cooper and T. Sueyoshi (1988b) "More on Breaking up Bell," Research Report CCS 590 -- Available as long as supplies last from W.W. Cooper, Graduate School of Business, University of Texas at Austin, Austin, Texas 78712-1174, USA.
- Cooper, W.W., L.M. Seiford and K. Tone (1999) *Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References and DEA-Solver Software* (Norwell, Mass.: Kluwer Academic Publishers).
- Cooper, W.W., L.M. Seiford and J. Zhu (2000a), "A Unified Additive Model Approach for Evaluating Inefficiency and Congestion," *Socio-Economic Planning Sciences* 34, pp. 1-26.

- Cooper, W.W., L.M. Seiford and J. Zhu (2000b) "Slacks and Congestion: A Response to the Comments of R. Färe and S. Grosskopf," *Socio-Economic Planning Sciences* (to appear).
- Cooper, W.W., R.G. Thompson and R.M. Thrall (1996) "Introduction: Extensions and New Developments in DEA," *The Annals of Operations Research* 66, pp. 3-45.
- Diewert, W.E. and C. Parkan (1983), "Linear Programming Tests Of Regularity Conditions for Production Functions," In W. Eichorn, R. Henn, K. Neumann and R.W. Shephard, (eds.) *Quantitative Studies on Production and Prices*, (Warburg-Wien: Physica Verlag) pp. 131-158.
- Färe, R. and S. Grosskopf (1998), "Congestion: A Note," *Socio-Economic Planning Sciences*, 32, pp. 21-23.
- Färe, R. and S. Grosskopf (1998), "Slacks and Congestion: A Response," *Socio-Economic Planning Sciences*, 34, pp. 35-50.
- Färe, R. and S. Grosskopf (2000), "Research Note: Decomposing Technical Efficiency with Care," *Management Science*, 46, pp. 167-168.
- Färe, R., S. Grosskopf and C.A.K. Lovell (1994), *Production Frontiers* (Cambridge: Cambridge University Press).
- Färe, R., S. Grosskopf and C.A.K. Lovell (1985), *The Measurement of Efficiency of Production* (Boston: Kluwer-Nijhoff Publishing Co.).
- Hanoch, G. and M. Rothschild (1972) "Testing the Assumptions of Production Theory: A Nonparametric Approach" *Journal of Political Economy* 80, 256-275.
- McDonald, J. (1996) "Note: A Problem with the Decomposition of Technical Inefficiency into Scale and Congestion Components," *Management Science* 42, pp. 473-474.
- Varian, H (1993), 3rd ed., *Micro-Economic Analysis* (New York: W.W. Norton & Co.).