

“IN THE JUNGLE”: TOWARDS A COMMON DOCUMENTATION STANDARD FOR CGE- BASED EXPERIMENTS

SEBASTIAN HESS

Institute of Agricultural Economics, Georg- August- University
37073 Göttingen, Germany

(May 2005)

Introduction: Transparency matters!

Economic science is quasi experimental rather than strict experimental (Button and Jongma, 1995). Applied economic models enable researchers to conduct experiments within a theoretical framework that is linked to real world data (van Tongeren, et al., 2001). The experimental framework of computable general equilibrium (CGE) models enables economists to control for certain factors that can hardly be controlled for under real world conditions. CGE models have moved out of economists' computing laboratories (Hertel, 1999). They have become integral part of the complex process of public decision making; economists, even if they would want to, cannot limit their simulation results to purely academic discussions. Instead, many CGE- based publications nowadays are designed more for a heterogeneous audience outside academics than for the peer review process. This public involvement in economic research imposes public responsibility on economists themselves and on the way in which they employ their methodological tools and communicate their results (Giersch, 1990).

Economic theory is concerned with the efficient use of resources. This should also apply to the information (e.g. simulation results) generated by its applied methods (Alston, et al., 1995). Using CGE models to analyze complex policies should therefore be done in the most efficient way with regard to the potential benefit that CGE- based information can unfold within society. One of the most fundamental obstacles to an efficient use of CGE- based information is the often discussed ‘black box’ criticism (Hertel, 2002, Panagariya and Duttagupta, 2001). In the public environment in which simulation results are used, transparency is a powerful remedy for the ‘black box’ criticism (Francois and Reinert, 1997). One of the most transparency enhancing approaches with regard to different economic methods is the establishment of systematic quantitative comparisons of CGE- based simulations across different publications (Stanley, 2001).

However, within many CGE- based publications, the descriptions of experimental settings such as model, data and parameters provide (for various reasons) only incomplete information about the whole CGE- based simulation, and almost never enable a reproduction of the results. Whenever the causality between simulation output and experimental design is hidden, the potential for asymmetric information (Akerlof, 1970) between “buyers” and “sellers” of simulation results is especially high. Therefore, analyzing the way how economists provide information about their experiments can be a step towards potential improvements of transparency and (thus) efficiency within the market for CGE- based information.

This paper is organized as follows: the next section outlines our research question in detail; next, we employ a literature survey in order to collect data on the ways in which

economists document their experiments. Based on this survey, we employ an econometric model in order to identify those factors that explain simulation results on average most significantly. In the last section we outline preliminary steps towards a transparency- enhancing documentation framework for CGE- based experiments.

Can Simulations reasonably be described by few important ‘Cornerstones’?

Simulations based on CGE models are scientific experiments that contribute to science through the provision of reproducible results. These results are the outcome of a given experimental setting. Although the causality between research question, model and results most often remains subject to interpretation, the fact that results can be reproduced provides an objective gain in information (see von Hayek, 1979). Although reproducibility in general is possible for the objective part of any CGE- based experiment, high transactions cost often limit reproducibility in practice and as a result the ‘black box’ character of models and experiments has commonly been criticized (Hertel, 2002, Panagariya and Duttagupta, 2001).

The GTAP network has combatted this ‘black box’ criticism through its open source policy concerning models and data. Openness is a key principle of GTAP (GTAP, 2005). However, there is no general, accepted framework that guides users of GTAP data or GTAP models on how to publish the settings of their experiments. The documentation of model specifications, applied shocks and simulated results does not follow systematic rules and thus makes it often difficult to compare related studies with regard to the causal

relationships that drive simulation results most heavily. Therefore, transparency remains limited even within and between publications that employ the GTAP modeling framework, but especially between GTAP and other CGE models.

The most transparent and comprehensive way to document CGE- based assessments would be to publish all data, the model's programming code and a list of all exogenous parameters as is standard in natural science (Science, 2005). However, for publications with CGE- based simulations this is, due to complexity and the demand of the public audience, most often not a feasible approach. At the same time, this would not be a convenient way for policy makers to access the nature of a simulation experiment.

Figure 1 outlines the optimization problem of appropriate documentation for any CGE- based publication: authors face various restrictions concerning the amount of information they can provide about the nature of their experiments. The provision of experimental details is associated with rising cost due to informational overload (figure 1): Since most model documentations have the size of books and folders rather than pages and brochures, it is hardly ever feasible to attach the complete description of one specific experiment to any CGE- based policy analysis. The author therefore has to bear the cost of intransparency (figure 1) and will decide on a selection of experimental details that may outline the most important 'cornerstones' of an experiment rather than enable precise reproduction of the results. In other words: the optimal amount of documented information about an experiment minimizes the sum of cost that is due to intransparency (i.e. forgiving information) and informational overload within a specific publication. Usually, the author of a CGE- based experiment approximates this optimum through the

selection of key information about the experiment that she/he considers to be appropriate and crucial for an interpretation of simulation results.

This implies that a better understanding of this key information (i.e. those elements that determine the simulation results most significantly) can potentially lead to a more efficient use of CGE- based information if the black box between “buyers” and “sellers” of CGE- based information is further reduced.

[Figure 1 here]

We can assume that most authors are aware of the ‘cornerstones’ within their experimental settings and therefore try to follow this optimization strategy (figure 1) already. A quantitative survey across CGE- based publications can therefore provide insight into the information that authors on average found to be the most crucial ones for an understanding of their experiment. Moreover, if simulation results can be linked to documented information, it might be possible to identify those ‘cornerstones’ econometrically: Variables that explain on average most of the variance within simulation results across studies provide on average most transparency for a given amount of information (figure 1). Equation (1) outlines a general model that identifies the relative importance of documented features within a simulation experiment:

$$(1) \quad Y_{r,n,e} = M_{r,n,e} + E_{r,n,e}$$

With Y = simulated change of output variable (e.g. welfare) in region (r) for study (n) in experiment (e); M is a vector of variables that authors have documented to be part of their model and may not vary across studies; E is a vector of variables that are unique to the simulation experiment, e.g. shock size, elasticities, tariffs, etc.

Improved knowledge about the explanatory ‘cornerstone’ variables that explain on average changes within the dependent variables best can guide authors of further simulation experiments on the question which information they should at least document along with a simulation experiment (as long as complete coverage of all experimental details is not possible). On the other hand, if authors decide to deviate from the average of documented key information and instead provide different details about their simulation experiments, their experiments itself will as well potentially be different from the observed average - a fact that may be a valuable piece of information especially for those members of the audience who rather focus on simulation results than on the experimental details.

In addition, the provision of more comparable information about the key factors that drive a simulation would enable better comparisons across publications and modeling frameworks. The GTAP network would be an ideal platform to further promote this research and establish a framework of key information that should be provided along with any simulation experiment in order to enable systematic comparisons among studies. Improved systematic comparisons of CGE models can lead to various transparency enhancing applications such as the establishment of a meta- analysis of CGE models or the development of a model response surface for CGE- based policy assessments. This paper intends to provide a preliminary starting point for this discussion.

Methods: How do Authors Currently Document their Experiments?

A Literature Survey of Trade Simulations under the Doha Round

CGE models are used by economists around the world: the institutions that frequently publish CGE-based analyses range from universities and large international organizations to national departments and private trade consultancies (GTAP, 2005). An appropriate literature sample therefore has to represent this range of institutions and should not exclude potential modeling applications through the sampling procedure. No unique database exists that would match our sample criteria, but we assume that “the internet” is the least biased database. Any sub-sample of this universal database is defined by the search process (e.g. the combination of keywords)¹. Through the combination of the search words “Doha”, “trade” and “GTAP” a manageable sample of 346 pdf. files could be defined². This broad specification of search words enables the inclusion of potentially very heterogeneous publications.

After eliminating duplicates and missing or erroneous links, 198 single publications³ have been identified (according to title and/or authorship). Of these, 86 contained tables that present simulation results (the remaining are mostly qualitative discussions of Doha trade issues which cite GTAP related work). These 86 studies again have been surveyed according to the following procedure:

For each publication, it has been investigated if and how the nature of the economic experiment has been documented. In general, several steps in the modeling process can be identified that frequently take place in almost every publication (figure 2). All steps

can have significant influence on the simulation results and a lack of documentation for any of these steps potentially reduces transparency. For instance, when a database is projected into a year in future the potential loss of transparency due to hidden assumptions can be especially high although the model itself might be documented in detail.

[Figure 2 here]

In order to capture the complexity of each experiment as outlined in figure 2, we establish a questionnaire that tracks the various aspects of CGE modeling and at the same time emphasizes the depth of each features' documentation by assigning a numerical rating. This questionnaire has been based on the knowledge of qualitative CGE comparisons and handbooks (Brockmeier, 2003, Francois and Reinert, 1997, Ginsburgh and Keyzer, 1997, Hertel, 2002, Lippoldt and Kowalski, 2003, Robinson, 1989, Scollay and Gilbert, 2000, Shoven and Whalley, 1984, UNCTAD, 2003, van Tongeren, et al., 2001).

More specifically, the survey framework that we employ via the questionnaire assumes that every CGE model exhibits the following aspects through one or the other specification: consumer side, producer side, markets that link both; trade that links markets between regions, and a time horizon within each experiment has to be interpreted; in addition, technical aspects (e.g. parameters, data, elasticities and computational issues).as well as assumptions about policy (in baseline and shocks).

The representation of policies as a major component of any CGE application is less pre-structured by any theoretical framework than technical descriptions of the average model or a widely used database. Although the evaluation of tariff changes is at the core of

classic CGE applications for trade policy analysis, nowadays CGE models are employed to analyze very heterogeneous policy changes (Hertel, 1999, Hertel, 1997). Beyond the “classic” trade policy implementation of price wedges into CGE models, the realistic representation of more complex phenomena is a major challenge for applied modeling (Brockmeier, 2003, Lips, 2004).

We started out with the rather general framework as outlined above and successively created new categories into which we have grouped the available information that has been extracted from the literature within the survey. This iterative approach has the advantage that no information is lost due to *a priori* omitted categories.

For any of the documented experimental features it has been assumed that four distinct levels of documentational detail exist. At the first level, a technical feature of a CGE model is generally mentioned to either be part of a simulation or not. This first level of documentation implies that the author states to have this specific component included into the analysis; however, the potential influence of this feature on the simulation results is not discussed. For example, the following statement is typical for the model description within many publications: “We employ a computable general equilibrium model with perfect competition and constant returns to scale. Our model is similar to the standard GTAP model described in Hertel (1997).” The survey questionnaire would extract the following information from this statement: the model belongs to the GTAP modeling framework; it has CRTS and perfect competition in all markets. All these aspects have been mentioned and thus each aspect receives a level 1 rating on the documentation rating scale.

The second level of documentary detail applies if a feature has been mentioned to be part of the model and at the same time the potential impact of this feature on the simulation results is discussed. For example the following statement would be rated at level 2: “Our model is comparative static in nature, this implies that the model is solved for one point in time, although the effects of the shock to the model have to be interpreted over a several years’ horizon.” We assume that documentation in this way contains significantly more information than in the first case because the reader is provided with a hint on how to interpret the result with respect to this specific modeling feature.

An even more detailed description of a simulation component would be the presentation of a graph, a table or an equation. We assume, for instance, that the general equation of a Constant Elasticity of Substitution Function (CES) provides the reader of a CGE- based experiment with more intuition about its role within the model than the verbal description of this equation does. Similarly, for instance the presentation of all regional and sectoral aggregations in tables is more comprehensive and easier to access than a, potentially incomplete, verbal description of regions and sectors in the text.

The highest level (4) of documentation is assigned to the presentation of any model component through its formulation in the specific software language (e.g. GAMS, GAUSS or GEMPACK). Clearly, the rating system that we have applied should not be understood as a hierarchical rating according to a “more details are better” criterion. The computational implementation or the presentation of a formula may not be efficient depending on the needs of the audience (compare figure 1). Therefore, our rating

system is not normative but captures the level of descriptive detail in a positive, numerical way.

In each of the charts presented in the next section (and the Appendix), the first bar presents the maximum number of rating points that can possibly be achieved according to the structure of our questionnaire. This maximum rating would be reached if all studies ($n = 86$ plus 2 studies that present simulations from two models; in total $n = 90$ studies) were to describe a specific feature by its implementation in software code (rating level 4; $n = 90 * 4 = 360$ rating points maximum). Clearly, this is neither useful nor realistic for most modeling components and publications.

Quantifying how economists document their experiments is not free of subjectivity; however, since the established categories are rather broad in focus and clearly distinct from each other, different evaluators on average would, according to our experience, agree on very similar ratings.

Results 1: Trends & Shortcomings of Current Experimental Documentation

Within our literature sample, more than 90% of the publications employ a version of the GTAP database, while roughly 49% (44 of 90) make direct use of the GTAP modeling framework (Figure 3). In other words: if, according to figure 3, 49% of all studies within the sample employ the GTAP modeling framework, one might expect certain GTAP features to be documented by an average of 44 publications within the sample. This is not the case as the figures 4 – 6 point out (figure 4-6 provide a summary of the survey results).

[Figure 3 here]

In figures 4-6 the ratio $\rho = \text{"total rating score from survey"} / \text{"number of observations"}$ yields the average level of documentational detail for each component (figures 4 – 6). This measure both provides an indicator for the effort that authors tend to devote on average towards the documentation of this feature and at the same time indicates to what extent the documentational depth varies between studies: at $\rho = 1$, all authors agree that it is appropriate to simply mention this feature (similar, for $\rho = 2;3;4$). On the other hand, if ρ deviates from straight values of $\rho = 1;2;3;4$ it can be observed that there is a heterogeneous way how authors decide to document this specific feature.

Figures 4-6 show that the ratio between number of observed studies that document a certain feature and the documentational depth for this specific feature varies widely from model component to model component. In general, no feature can be observed to be documented on an average level of 3 or 4. Documentation for the production- and the consumption side within the model appears to be rather homogeneous (figure 4 and 5). At the same time the documentation of data, time horizon and parameter issues is, according to figure 6, very heterogeneous because of uneven values for ρ .

This can either imply that one specific feature has a widely differing impact on the results of the experiments involved and therefore is documented with differing intensity, or, that authors have very different strategies how to solve the optimization problem outlined in section 2 (figure 1).

The rather straight values for p with regard to model components (figures 4 and 5) can be an indication for the fact that especially users of the GTAP modeling framework show a tendency to follow the model documentation that is stored somewhere else (e.g. Hertel, 1997) and which allows to limit experimental descriptions to rather general statements. At the same time, the number of studies show a tendency for general features to be well below sample size. This implies that most features are mentioned only by very few of the studies in focus. For instance, less than one third of all studies mention their closure rules. Certain aspects of capital accumulation or specific policy shocks that have been implemented appear to be the most frequently mentioned aspects in this regard. The value for p however indicates that those authors who discuss their closure tend to do this in a very detailed way.

The fact that the number of observations is significantly lower than the sample size (i.e. many authors do not report on a certain feature at all) can be for two reasons: Either, within a publication a certain feature is not used and thus the authors have little reason to discuss it, or, on the other hand, the authors consider the documentation of this feature to be of minor importance for an interpretation of their simulation results, given the publication- specific restrictions outlined in section 2. This dichotomy of reasons that significantly reduces transparency within and between publications because no one besides the team of modeling authors can know precisely for each feature whether it is part of the model or just happens to get omitted from the documentation. For the purpose of any form of (quantitative) model comparisons, this introduces a measurement error that will be briefly discussed along with our econometric analysis in the next section.

[Figure 4 & 5 here]

If figure 6 indicates that the documentation of components outside the actual modeling framework tends to be rather *ad hoc*, this does not hold true for the database. Although the GTAP- database in its various versions is obviously standard within the CGE community, authors devote comparatively much effort to the description of this database and the aggregation they have chosen. The aggregation scheme is mostly presented through tables. In addition, about half of the publications within the sample somehow mention the importance of various exogenous and endogenous parameters for their simulation. If own estimates are presented these are usually documented in detail.

[Figure 6 here]

About 1/3 of the publications include additional tariff data into the base data in order to produce a more realistic data environment for the specific policy in focus. The ongoing improvement of the GTAP database with regard to the inclusion of applied tariffs is well represented within the literature sample (e.g. Fontagné, et al., 2002, UNCTAD, 2003). More than one third of all studies within the sample make use of modified additional tariff data (figure 6). The level of documentation detail tends to be high for this feature. In other words, authors generally tend to discuss their tariffs in detail.

It occurs quite often that the database is projected into a different base year, although it cannot be considered standard under the comparative- static framework that most studies within our sample are using. Data base years range from 1995 to 2002 and projections range from 1995 to 2025 (includes dynamic scenarios). In this context the assumption of China's WTO accession and as a result of that the implementation of pro-trade policies

are frequently mentioned. Assumptions about growth rates are however reported less frequently. An average aggregation of 10 to 15 regions and 15 to 18 sectors is common, although the variance is quite large.

The “standard CGE model” that most authors in our survey refer to is often, but not always, synonymously used as a description for the standard GTAP model as described in Hertel (1997). Most authors describe this model as CRTS, perfect competition in all markets and comparative static in nature. Some authors add a description of the capital market closure and frequently mention the regional differentiation of goods (Armington assumption (Armington, 1969)) when they describe this model.

Researchers often devote a significant amount of space within any CGE- based publication to the description of assumptions that abstract from a real world policy proposal in order to enable a more or less realistic representation of specific policies within their model (Brockmeier, 2003, Lips, 2004). On the other hand, some publications design the applied policy shocks in a rather simplified manner that can *a priori* be managed by the existing model, for example the complete elimination of all tariffs in all markets world wide (UNCTAD, 2003, Vanzetti and Peters, 2003). Therefore, the variance within the way in which researchers present and document the policy part of their analysis to some extent “naturally” depends on the policies to be analyzed; no general standard of documentation could eliminate this variance in a useful way. Policies which are comparatively complicated to implement into a CGE model (e.g. Swiss formula, CAP sugar market reform) tend to be documented in less detail.

In summary, our survey has quantified the degree to which authors document experimental details of their simulation experiments. As a result, no clear picture can be established that would identify systematic patterns about the way how experimental settings are described. A tendency can be observed to frequently report on rather general aspects of the model, while technical details, such as functional forms, shocks, closure rules and parameters tend to be described in a way that varies strongly between publications. We have tried to econometrically link external influences on the publicational level to the overall degree of experimental detail. We failed to establish significant causal relationships, which is an indicator for the fact that authors rather intuitively decide which ‘cornerstones’ may provide most transparency about their experiments and therefore deserve more or less intense documentation.

Results 2: Which of the documented modeling features create most transparency?

The previous analysis has shown that on average authors provide information about their experimental settings on a wide range of details; these details vary significantly between studies. In this section we investigate whether this information is a meaningful way for an interpretation of the results and which of the documented details are on average the most important ones.

We use the same dataset as described in the previous section, except the rating points of documentation detail that have been transformed into binary information (dummy variables). For example, a study that mentions the use of the Armington assumption (rated at level 1) will now be described with the Armington dummy (at value 1). A

different study that describes the implementation of the Armington assumption in detail (e.g. at rating level 3) will also be represented by the Armington dummy (value 1). All variables within the dataset have been rescaled according to this procedure because we assume that the depth of experimental documentation does not influence the functioning of a feature within the model. In order to save observations, the missing values of numeric variables have been replaced by the sample mean. In the case of binary variables, it has been assumed that missing information implies that a specific model component is not employed (variable takes the value 0).

The latter assumption does not distinguish between experimental details that have not been mentioned, although they are part of the model, and those that have not been mentioned within the documentation because they are not part of the simulation exercise. This limits the interpretation of the results: Two identical models from different publications are treated differently within the dataset according to the information that the authors have provided. This type of measurement error within the explanatory variables can be described in the following way (Hsiao, 2003, Wooldridge, 2003):

$$x_i = x_i^* + \tau_i$$

; x is the true variable, x^* is observed
; τ = measurement error due to missing documentation
; $i = 1, \dots, N$ observations
: $x^*, x, \tau \in \{0,1\}$

There are several ways to control for measurement error within explanatory variables (e.g. instrumental variables or “within- family” comparisons (Hertz, 2003)). However, most of these methods require additional information (compare e.g. Li and Luo, 2004) that could be obtained in the case of our survey through the request of detailed

information from the authors. Since we wish to work on the basis of documented information we proceed with the assumption that authors on average document the most relevant features of their models and therefore ignore the attenuation bias (Wooldridge, 2003) that may arise from measurement (documentation) errors.

The largest group of studies within the sample (36 out of 86) report absolute changes in equivalent variation (EV) for each region (r) as a measure of the simulated effects. The remaining studies focus on percentage changes in EV relative to the base level, chose the change in GDP as the variable of interest, or focus on completely different variables. In general these welfare measures cannot be directly compared because they are either based on different welfare concepts (Varian, 1992) or they are derived from different variables within a CGE model (Hertel, 2004). Some authors focus on the change in GDP or other output variables such as quantity changes or trade volume changes. If, in addition, the results for each region are expressed in relative terms, very different interpretations for the output of a simulation experiment become necessary. Since the largest sub sample of publications reports absolute changes in EV, this subset has therefore been chosen for our regression analysis.

Based on the general model specified in equation (1) we fit a linear regression model (2) that explains simulated welfare changes at the regional level⁴ as a function of applied shocks and documented characteristics of the CGE models within the survey.

Our general model is of the following form:

$$\begin{aligned}
(2) \quad Y_{r,n,e} &= \beta_0 + \beta_1 X_{1\ r,n,e} + \beta_2 X_{2\ r,n,e} + \varepsilon && ; r = 1, \dots, 2827 \\
&&& ; n = 1, \dots, 36 \\
&&& ; e = 1, \dots, 168
\end{aligned}$$

With Y = simulated welfare change (EV) in region (r) for study (n) in experiment (e); X_1 is a vector of binary variables that describes whether a study employs a specific feature or not. X_2 is a vector of variables that takes on numeric values, e.g. base year, or shock size. For ε we assume $\varepsilon \sim N(0, \sigma^2)$.

Table 1 presents the results of our linear model (OLS estimator). The Breusch-Pagan Test rejects the hypothesis of homoskedasticity; we have therefore calculated robust standard errors for the coefficients (table 1; table 2 in the Appenix provides a descriptive summary of dependent and explanatory variables). The estimation results for the dependent variable are difficult to interpret: a given simulated change of welfare can imply a large gain to a small economy or a comparatively small gain to a large economy. Our regression model cannot distinguish between these differences. The inclusion of aggregated GDP and population figures for each region within the dataset increases the overall fit of the model significantly, however, this information would complement the set of variables documented within the studies by information from outside and has therefore been omitted.

We have tested all variables that have been obtained from the survey according to an iterative inclusion process during which we have selected according to the highest level of significance. Since the dependent variable is measured as equivalent variation in millions of US Dollars, the estimated coefficients must be interpreted in the same way:

Any coefficient is the deviation of simulated welfare change in region (r) from the baseline in absolute terms. Many estimated coefficients bear a negative sign. This can be attributed to the baseline regression scenario: the hypothetical observation behind this regression model takes on the value zero for all variables within the vector X_1 . This is not a realistic scenario because none of the model components that have been surveyed are present in this baseline model. Due to the high estimate for the intercept, all other dummy variables that take the value 1 are downwards- shifted intercepts relative to this baseline intercept.

Although all variables (except the dummy for the SMART model) that have been included into the model are significant, due to the large sample size the commonly applied 5% probability level (95% significance) is not strict enough. For sample sizes of more than 1000 observations a significance level of 97.5 or 99 should be applied (Hendry and Krolzig, 2001). We have therefore reported the probability levels to the fifth digit after the comma.

[Table 1 here]

The estimated coefficients for several key variables are broadly in range with qualitative knowledge about the effects of these features within simulation results. The estimated sign of the projection year ($\beta_{\text{Projection to Year}} = 5726$) is positive. Projection years far into the future are mainly the result of dynamic simulations. Dynamic models are known to produce larger welfare gains than comparative static models on average (UNCTAD, 2003); the estimated effect is also positive ($\beta_{\text{Dynamic?}} = 55341$).

Authors who include their own elasticity estimates ($\beta_{\text{Own Elast. Estimates incl?}} = 233823$) or double their Armington elasticities ($\beta_{\text{Armington doubled ?}} = 117629$) within their publications show higher welfare estimates on average; the assumption that China and Taiwan join the WTO before 2005 and gain full access to developed country markets is also associated with welfare gains above average ($\beta_{\text{Med. Term Policy optimistic?}} = 122085$). Our results point to the fact that studies with those modifications show comparatively large simulated welfare changes.

The coefficients for $\beta_{\text{Inc>Returns to Scale? (Y/N)}} = -73888$, $\beta_{\text{New Growth Features?}} = -156083$ and $\beta_{\text{Imperfect Competition?}} = -237612$ exhibit rather large negative signs, which is not expected.

Often, these features are attributed to rather positive welfare effects for specific regions, meanwhile it is possible that many other regions within an experiment experience welfare losses due to these features (Elbehri and Hertel, 2004, Francois and Roland-Holst, 1997).

The absolute value of the estimated coefficients as presented in table 1 can, due to attenuation bias, not be regarded reliable and should not be subject to rigorous interpretation and analysis. Due to missing documentation, the results will most likely be biased towards zero. The very high estimate (in absolute terms) for the intercept ($\beta_{\text{Intercept}} = -11687320$) suggests that the average region within the sample experiences a welfare loss of this amount if all variables take the value zero. Since this is not realistic, it can be concluded, that a significant share of the variance within the dependent variable is estimated to be an output of the ‘all variables zero model’. It is quite clear that such a model cannot produce any output and therefore the results are most likely attenuated towards zero.

Since our literature sample addresses a very broad focus of CGE applications, our econometric model nevertheless demonstrates that comparatively few significant variables can be identified which are frequently reported by authors and prove to explain a certain share of the variance within the simulation output. In addition, large coefficients have been estimated for parameters that are not covered by the standard model documentation (e.g. (Hertel, 1997) because these features are unique to each experiment. Our descriptive analysis has shown that especially for these elements of a simulation experiment the documentational level within the sample appears to be scattered and fragmentary.

The estimated coefficients can overall be related to qualitative knowledge about the impact these features have on the simulation results. This opens the scope towards a new strand of research: If attenuation bias would be controlled and certain other econometric issues such as the hierarchical structure of the dataset would be taken into account, the impact of experimental ‘cornerstones’ could be estimated across experiments and across publications. The estimated coefficients would present guidance on which experimental features are the most crucial ones for a given set of applications and should therefore at least be documented or should be subject to a systematic sensitivity analysis within the model. Furthermore, the inclusion of variables about the studies enables a meta-analysis (Stanley, 2001) that relates simulated output to experimental features and factors that are external to the experiments.

Discussion

The previous analysis has shown that various ‘cornerstones’ of general experimental documentation can be identified: The information that authors’ document along with their experimental settings can be linked econometrically to the simulation results. The estimated coefficients of many significant variables are broadly in range with qualitative knowledge about the influence of these specific features on simulation results. Authors manage to provide their audience with reasonable key information on the nature of their experimental settings in such a way that causal relation between this information and the simulation results can be established.

The large part of variance that remains unexplained (according to R^2) and the high level of the intercept suggest that important variables may not have been documented at all in some publications. This might be addressed using instrumental variables which correct for the attenuation bias (Li and Luo, 2004).

Figure 1 in the section 2 has outlined the problem that authors face with regard to the presentation of a simulation experiment: Depending on various factors such as the specific demand of the audience and the nature of the publication, it will in most cases not be feasible to fully document the experimental settings.

As a remedy to this, it is according to the results of our analysis possible to identify certain ‘cornerstones’ that happen to get documented in many simulation studies and at the same time prove to explain simulation output significantly. Based on this, some guidelines emerge that enable more transparency within model documentations without

additionally restricting authors. These guidelines may be a starting point for the development of documentational conventions within the community of CGE modelers and can be outlined as follows:

For any publication, authors would first decide which of the following three documentation categories to choose:

- **Type 1:** No documentation is provided.
- **Type 2:** Comprehensive documentation allows for systematic comparisons.
- **Type 3:** Full documentation enables the precise reproduction of simulation results and allows tracking of all causal relationships within the experiment.

Documentation styles according to ‘type 1’ and ‘type 3’ are very clear in their interpretation: The results of our survey suggest that there are certain conditions under which authors and institutions find it appropriate to not provide any details about the experimental settings of their simulations and rather focus on the interpretation of results.

In this case, for instance, researchers who are interested may request more details on the nature of an experiment directly from the authors. Alternatively, a hyperlink could provide access to ‘type 3’ documentation for any CGE- based simulation. In fact, several publications within the sample show hyperlinks to a more detailed documentation or provide contact information from which more specific information can be requested.

Almost all studies within our sample describe their experiments with some detail in order to provide ‘semi- insiders’ with a brief characterization of what has been simulated. All these efforts can be summarized under ‘type 2 documentation’. Experimental documentation according to ‘type 2’ could be based on any widely known modeling

approach such as the standard CGE model as described in (Hertel, 1997): Users of GTAP often cite the GTAP model as “the standard CGE model”. These authors could extend their documentation by explicitly mentioning all modifications they made to this “standard model”, for example “we use the standard GTAP model version xy. ALL modifications we made are listed below:” Furthermore, all authors should agree to report simulation output in terms of welfare change in EV or GDP as the most widely documented output variables- regardless the specific output variable they focus on within their publication. Information about regional and sectoral aggregation as well as information on additional other data and/or parameters that may have been included would provide a comprehensive proxy to the experiment that is parsimonious in space. This should be included in every publication and would ensure a maximum of possible transparency, as our econometric approach has demonstrated.

The ‘type 2’ documentation approach is flexible concerning new modeling innovations: Whenever the “standard” GTAP model moves on, the latest version serves as the reference point. In principle, even non- GTAP users could follow this strategy by expressing their model relative to a version of the standard GTAP model: “We employ a CGE model that is comparable to the GTAP model version xy, with the following exceptions: our agricultural sector exhibits decreasing returns to scale, ... ”. As a general rule for a more efficient documentation standard according to “type 2” documentation the following can be summarized: Authors should

- Firstly, choose an established and comprehensively documented modeling framework that is closest to the particular model in use. They should express ALL

changes relative to this reference model. These changes do not have to cover every detail; rather, all changes should be expressed in terms of general categories (in order to minimize documentation).

- Secondly, mention the database in use, document the chosen aggregation and describe in a general way ALL changes and additions that have been made.

Since the average documentation of simulation experiments presents rather general facts, it will be more transparency enhancing for a single publication to have information on all features of the model available in a general way rather than to be provided with exemplary details at the expense of a comprehensive general view. Therefore:

- Mention all features first in the most general way and then successively add more details where it is most important for the experiment in focus.
- No matter which output variable is in focus of the specific analysis, a widely used welfare measure such as EV or Δ GDP should be mentioned for each experiment at least at the regional level.

Some publications within our sample contain a statement similar to the following: “... we have chosen elasticities that are more appropriate for the purpose of the analysis than the ones in the standard model.” This statement can leave the reader with a very inefficient ‘black box’ frustration. Instead, a statement such as the following, though far from providing comprehensive coverage, would not require any more amount of space, yet can provide a lot more inside into the nature of the experiment: “...we have chosen to modify

the following elasticities ..., on average, the values we use are three times higher than the ones in the standard GTAP model, version xy”.

In addition to these guidelines, authors would be free to discuss any new feature they have added to their model in much more detail; in addition, the context of the remaining model would be more transparent. Furthermore, authors are still free to post their full documentation on the internet or send it out on request. Journals with peer review could adopt these conventions; the space dedicated to model descriptions could generally be used more efficiently.

The regression model we use is a clear starting point. Within the sample, it has been the best possible fit based on observed model documentations. The method can easily be extended towards a systematic model comparison or a meta- analysis (Stanley, 2001) if the potentially existing attenuation bias would be eliminated and the regression model would include flexible functional forms and account for the hierarchical structure of the dataset (Goldstein, 1995). With more comparable and easier accessible documentation on experimental settings, it would for instance be possible to compile datasets of studies that all employ a specific feature of interest. Based on this, an econometric sensitivity analysis across studies and the development of a model response surface across various modeling approaches would become possible.

Conclusions

Although reproducibility and full documentation is not feasible for most CGE- based publications, it is necessary to achieve a maximum amount of transparency within and across CGE- based simulation experiments through the documentation of key elements ('cornerstones') that heavily influence most simulation results. Our analysis has shown that based on a randomly drawn literature sample of contemporary CGE applications, such key elements can be identified econometrically across a heterogeneous set of publications, although we cannot completely rule out that some additional key factors may not have been documented by any of the studies within our sample and therefore remain hidden. Our econometric results strongly support the view that insight into the model alone is not sufficient in order to enable an interpretation of simulation results. Instead, information about modifications to data, assumptions about medium term policies and the nature of applied policy shocks is of potentially equal importance. Our descriptive survey results suggest that especially these components of a CGE- based experiment tend to be documented in a very heterogeneous and ad hoc manner. Each author faces the problem to ensure a maximum amount of transparency for a large number of factors that drive the results of each specific experiment. Therefore, we have, based on the results of our analysis, started to outline a systematic framework of guidelines on how to present the nature of a simulation experiment in a more efficient and transparent way. It should be clear that this can only be a starting point for further discussions within the community. In addition, our econometric analysis points towards

the fact that further research based on ‘cornerstone’ information about CGE- based publications is possible: If more focus is placed on the information about institutional settings and authorship, for instance, a meta- analysis of CGE models and partial models would potentially lead to a significant increase in transparency within trade related policy simulations. Including numerical information about different parameters would enable a model response surface.

Combating the old ‘black box’ criticism could further strengthen the role of applied economic modeling for policy advice. The econometric overview that can be generated across modeling applications can reduce the potential for simulations that depend crucially on questionable key assumptions and therefore cause confusion and lose credibility among the audience (Kehoe, 2002, Panagariya and Duttagupta, 2001). Thus the role of applied quantitative policy analysis would further be strengthened and could support public discussions and opinions with better quantitative facts. Public economists would find it easier to defend their arguments (Giersch, 1990) and, finally, a more efficient use of model- generated information could even help to speed up trade negotiations itself. A systematic documentation of experimental settings is a starting point for this.

¹ For instance, “Google” as the most often used internet search engine traces up to 1000 of the most relevant hits (hits beyond 1000 are estimated but cannot be accessed via the search engine web site). We assume that all recent papers of potential relevance which are electronically published can be found as .pdf files. Details about the search process can be requested from the author.

² According to our experience, “GTAP” is a useful keyword for an internet search that matches the goal of this specific analysis. The alternative keywords “CGE” or “model” have lead to the inclusion of many studies that only cite results from CGE- based experiments without presenting own estimates.

³ A complete list of all publications within the sample is available on request.

⁴ Authors frequently report summaries of regional effects such as “Africa total” or “World total”. Since these figures are usually weighted sums of other reported regions, all these aggregations have been deleted from the dataset in order to avoid an overrepresentation of large economies. “Rest of the World” aggregations (ROW) have not been removed from the dataset although they tend to be heterogeneous groupings of various regions outside the policy focus; the reported welfare effects for ROWs may not bear much of an economic message either. However, ROW aggregations have not been removed from the dataset in order to present for each simulation a complete set of regions.

References

- Akerlof, G. "The market for "lemons": quality uncertainty and the market mechanism." *The quarterly journal of economics, MIT Press* 84, no. 3(1970): 488-500.
- Alston, J. M., G. W. Norton, and P. G. Pardey. *Science under scarcity : principles and practice for agricultural research evaluation and priority setting*. Ithaca [u.a.]: Cornell Univ. Press, 1995.
- Armington, P. S. "A theory of demand for products distinguished by place of production." *IMF Staff Papers* 16(1969): 159-177.
- Brockmeier, M. *Ökonomische Auswirkungen der EU- Osterweiterung auf den Agrar- und Ernährungssektor*. Agrarökonomische Studien. Kiel: Wissenschaftsverlag Vauck, 2003.
- Button, K. J., and S. M. Jongma (1995) *Meta-analysis Methodologies and Microeconomics*. Amsterdam.
- Elbehri, A., and T. Hertel. "A comparative Analysis of the EU- Morocco FTA vs. Multilateral Liberalization." *GTAP Working Paper No. 31* (2004).
- Fontagné, L., J.-L. Guérin, and S. Jean (2002) *Multilateral Trade Liberalization: Scenarios for the New Round and Assessment*. Brussels.
- Francois, J., and K. A. Reinert (1997) *Applied Methods for Trade Policy Analysis- A Handbook*, Cambridge University Press.
- Francois, J., and D. Roland-Holst (1997) *Scale Economies and Imperfect Competition*, ed. J. Francois, and A. Reinert. Cambridge University Press.
- Giersch, H. (1990) *On being a Public Economist- Lecture held at the prize-awarding ceremony of the Paolo Baffi International Prize for Economics 1989 with a Laudation by Edmond Malinvaud*. Kiel.
- Ginsburgh, V., and M. Keyzer. *The Structure of applied general Equilibrium Models*. Cambridge, Massachussetts & London, England: The MIT Press, 1997.
- Goldstein, H. *Multilevel Statistical Models*. Kendall's Library of Statistics 3. London: Arnold, London- Sydney- Auckland, 1995.
- GTAP (2005) *Mission, Core Values and Vision*, vol. 2005, The Center for Global Trade Analysis, <http://www.gtap.agecon.purdue.edu/about/center.asp#vision>
- Hendry, D., and H.-M. Krolzig. *Automatic Econometric Model Selection Using PcGets 1.0*. London: Timberlake Consultants, 2001.
- Hertel, T. W. (2002) *Applied General Equilibrium Analysis of Agricultural And Ressource Policies*, ed. B. Gardener, and G. Rausser, vol. 2, Elsevier Science B.V.
- Hertel, T. W. "Future Directions in Global Trade Analysis." *Department of Agricultural Economics, Purdue University* (1999).
- Hertel, T. W. (1997) *Global Trade Analysis- Modelling and Applications*, Cambridge University Press.
- Hertel, T. W. (2004) *GTAP: Frequently asked Questions*, http://www.gtap.agecon.purdue.edu/resources/faqs/faqs_display.asp?F_ID=92
See also GTAP technical paper no 5 and no 20, vol. .
- Hertz, T. "Upward Bias in the Estimated Returns to Education: Evidence from South Africa." *The American Economic Review* 93, no. 4(2003): 1354-1368.

- Hsiao, C. *Analysis of Panel Data*. Cambridge, U.K.: Cambridge University Press, 2003.
- Kehoe, T. J. "An Evaluation of the Performance of Applied General Equilibrium Models of the Impact of NAFTA." *Paper Presented at the 5th Conference on Global Economic Analysis* www.econ.umn.edu/~tkehoe(2002).
- Li, H., and Y. Luo. "Reporting Errors, Ability Heterogeneity, and Returns to Schooling in China." *Pacific Economic Review* 9, no. 3(2004): 191-207.
- Lippoldt, D., and P. Kowalski. "The Doha Development Agenda: Welfare Gains from Further Multilateral Trade Liberalization with Respect to Tariffs." OECD.
- Lips, M. "The CAP Mid Term Review and the WTO Doha Round: analyses for the Netherlands, EU and accession countries." Agricultural Economics Research Institute (LEI).
- Panagariya, A., and R. Duttagupta (2001) The 'gains' from preferential trade liberalization in the CGE models, where do they come from?, ed. S. Lahiri, pp. S. 39-60.
- Robinson, S. (1989) *Multisectoral Models*, ed. H. Chenery, and T. N. Srinivasan, Elsevier.
- Science (2005) *Preparing Your Supporting Online Material*, vol. 2005, Science Magazine, http://www.sciencemag.org/feature/contribinfo/prep/prep_online.shtml.
- Scollay, R., and J. Gilbert. "Measuring the Gains from APEC Trade Liberalization: An Overview of CGE Assessments." *The World Economy* 23, no. 2(2000): 175-197.
- Shoven, J. B., and J. Whalley. "Applied General- Equilibrium Models of Taxation and International Trade: An Introduction and Survey." *Journal of Economic Literature* 22, no. September(1984): 1007-1051.
- Stanley, T. D. "Wheat from Chaff: Meta-Analysis As Quantitative Literature Review." *Journal of Economic Perspectives* 15, no. 3(2001): 131-150.
- UNCTAD. "Back to Basics: Market Issues in the Doha Agenda." Authors: S. Laird, L. Cernat, A. Turrini.(2003), Geneva.
- van Tongeren, F., H. van Meijl, and Y. Surry. "Global Models Applied to Agricultural and Trade Policies: A Review and Assessment." *Agricultural Economics* 26(2001): 149-172.
- Vanzetti, D., and R. Peters (2003) *Making Sense of Agricultural Trade Policy Reform*. Durban, South Africa, pp. 875-893.
- Varian, H. R. *Microeconomic Analysis, Third Edition*. New York/London: Norton, 1992.
- von Hayek, F. A. *Missbrauch und Verfall der Vernunft*. Salzburg, Fotomechanischer Nachdruck der 1. Auflage von 1959 (Verlag Knapp, Frankfurt/Main): Verlag Wolfgang Neugebauer, 1979.
- Wooldridge, J. M. *Introductory Econometrics*: Thomson, South-Western, 2003.

Appendix:

Figure 1: Pre selected documentation leads to a more efficient use of information.
Source: Own.

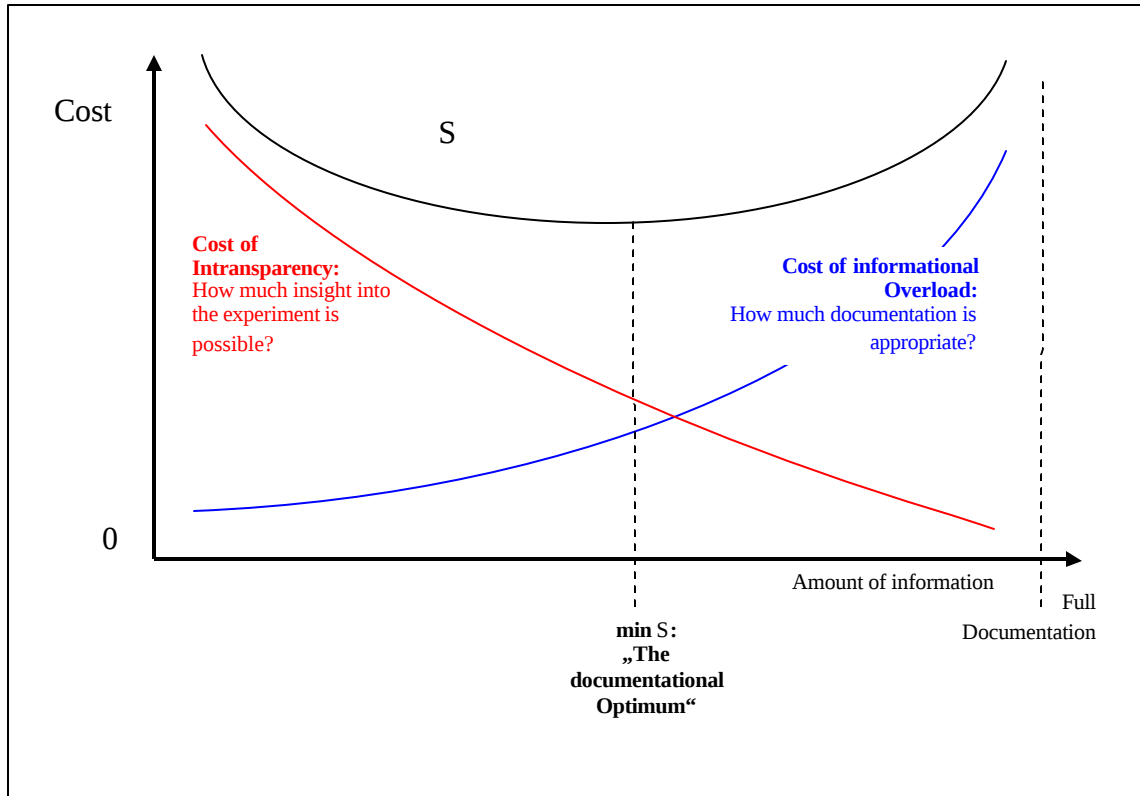


Figure 2: Each experiment is a combination of existing components and study specific modifications. Source: Own, based on Lips (2004).

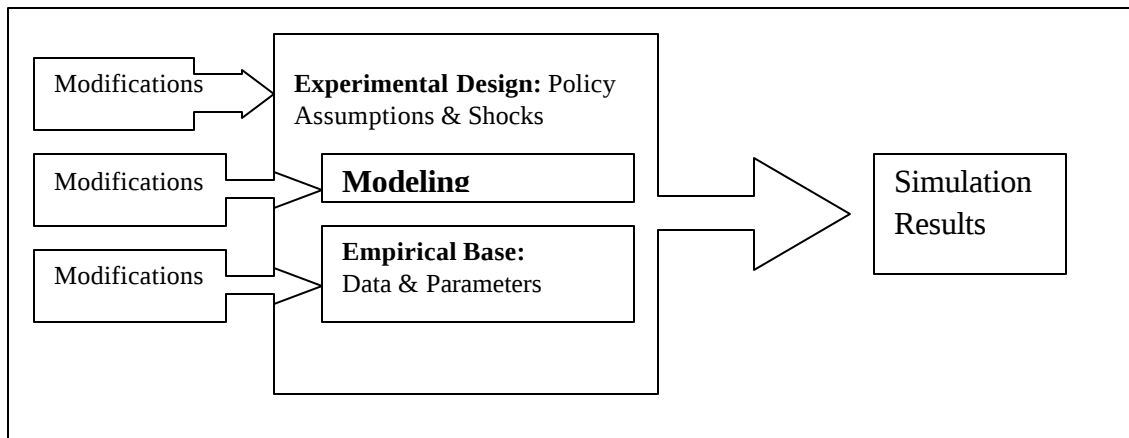


Figure 3: Modeling frameworks within the sample. Source: Survey Results.

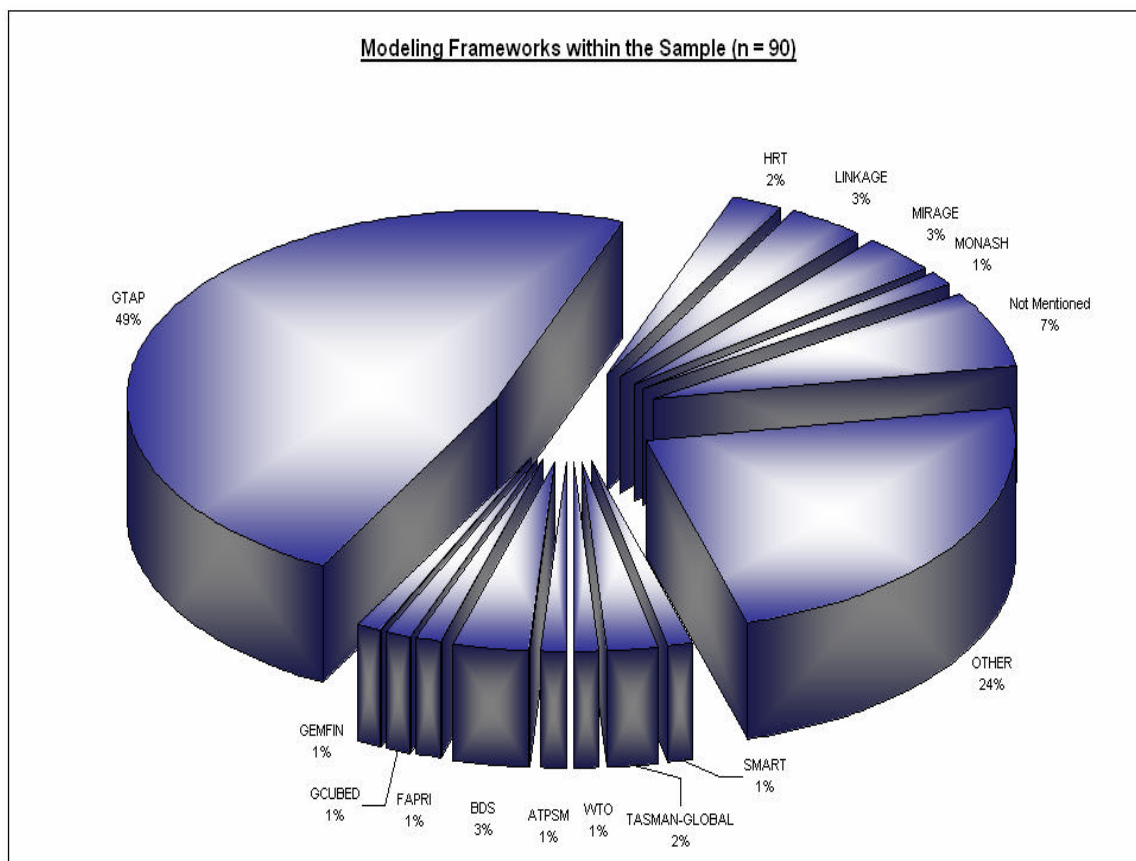


Figure 4: Documentation of the production side. Source: Survey Results.

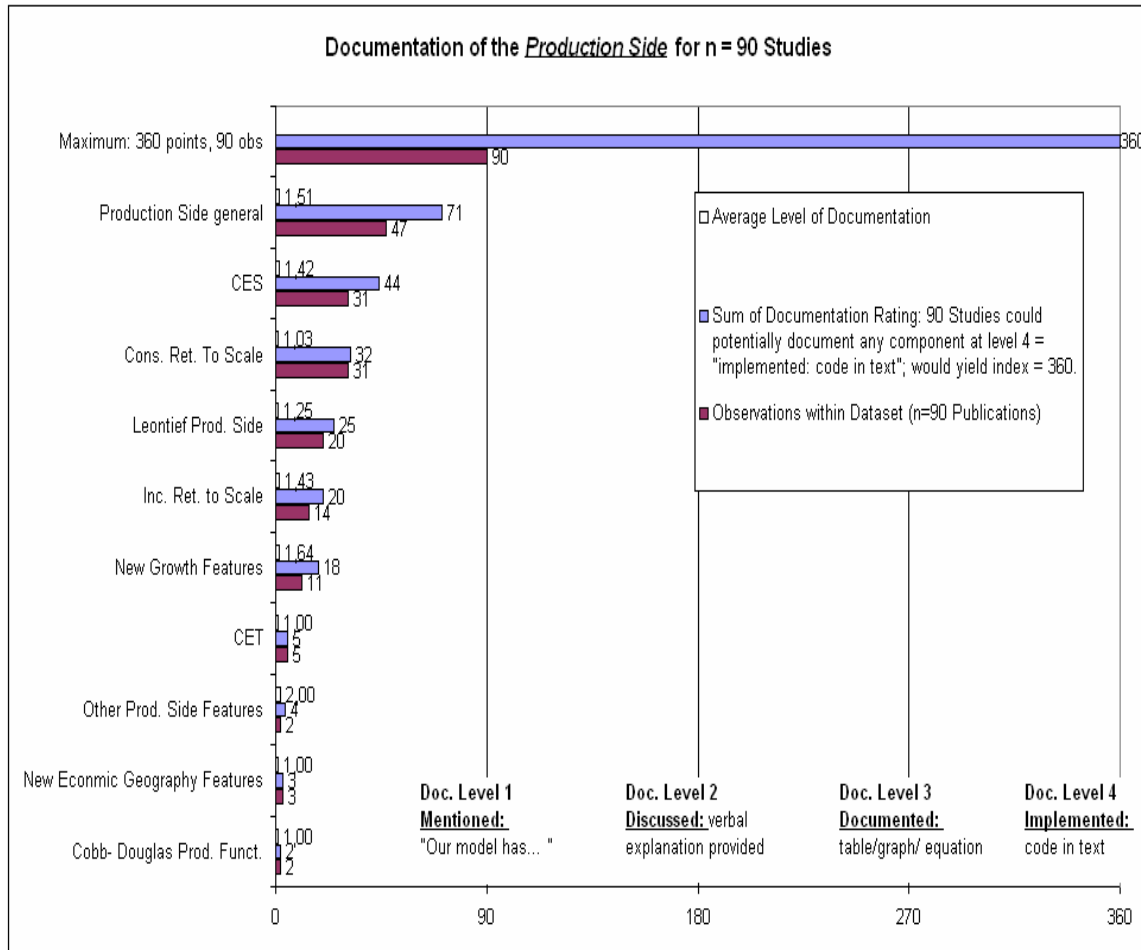


Figure 5: Documentation of the consumption side. Source: Survey results.

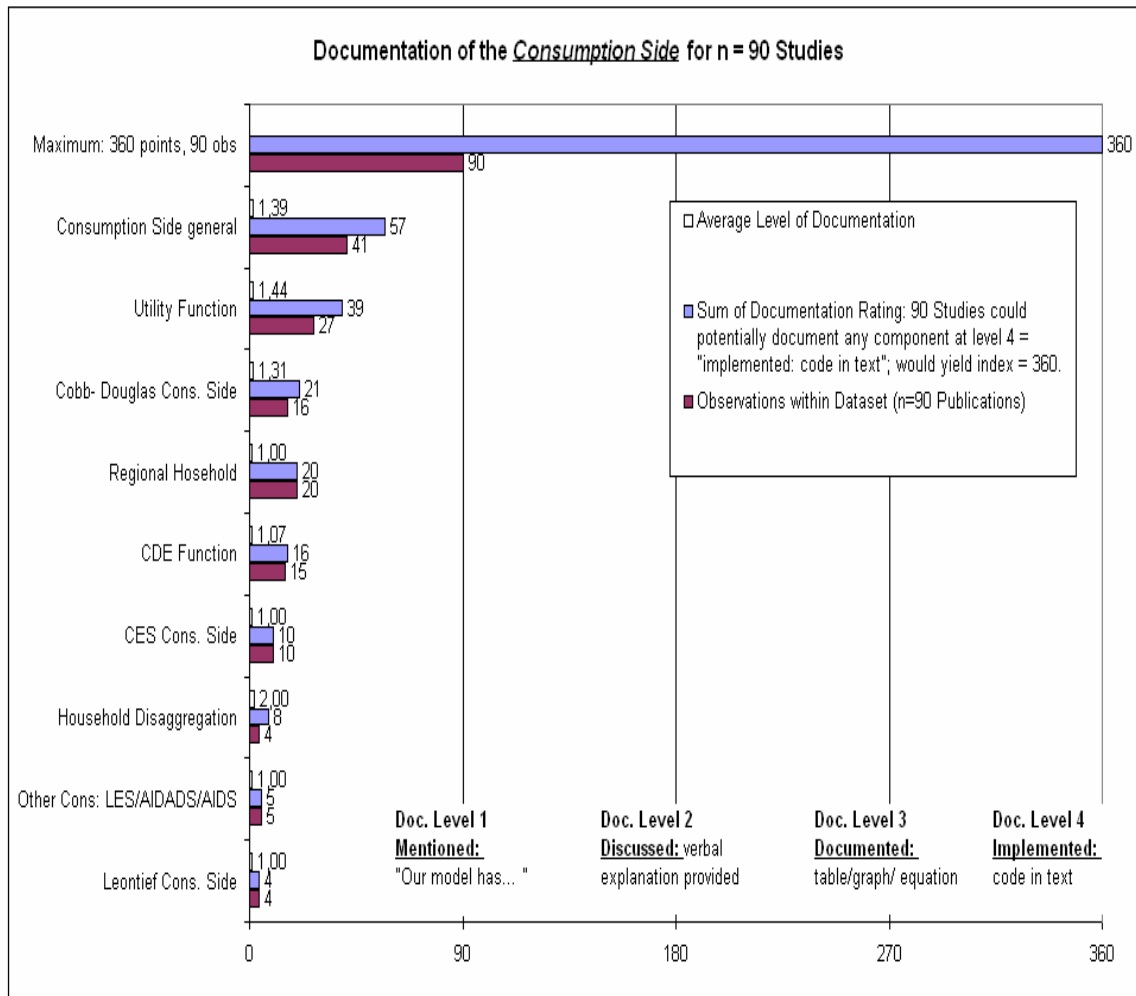


Figure 6: Documentation of database, time & parameters. Source: Survey results.

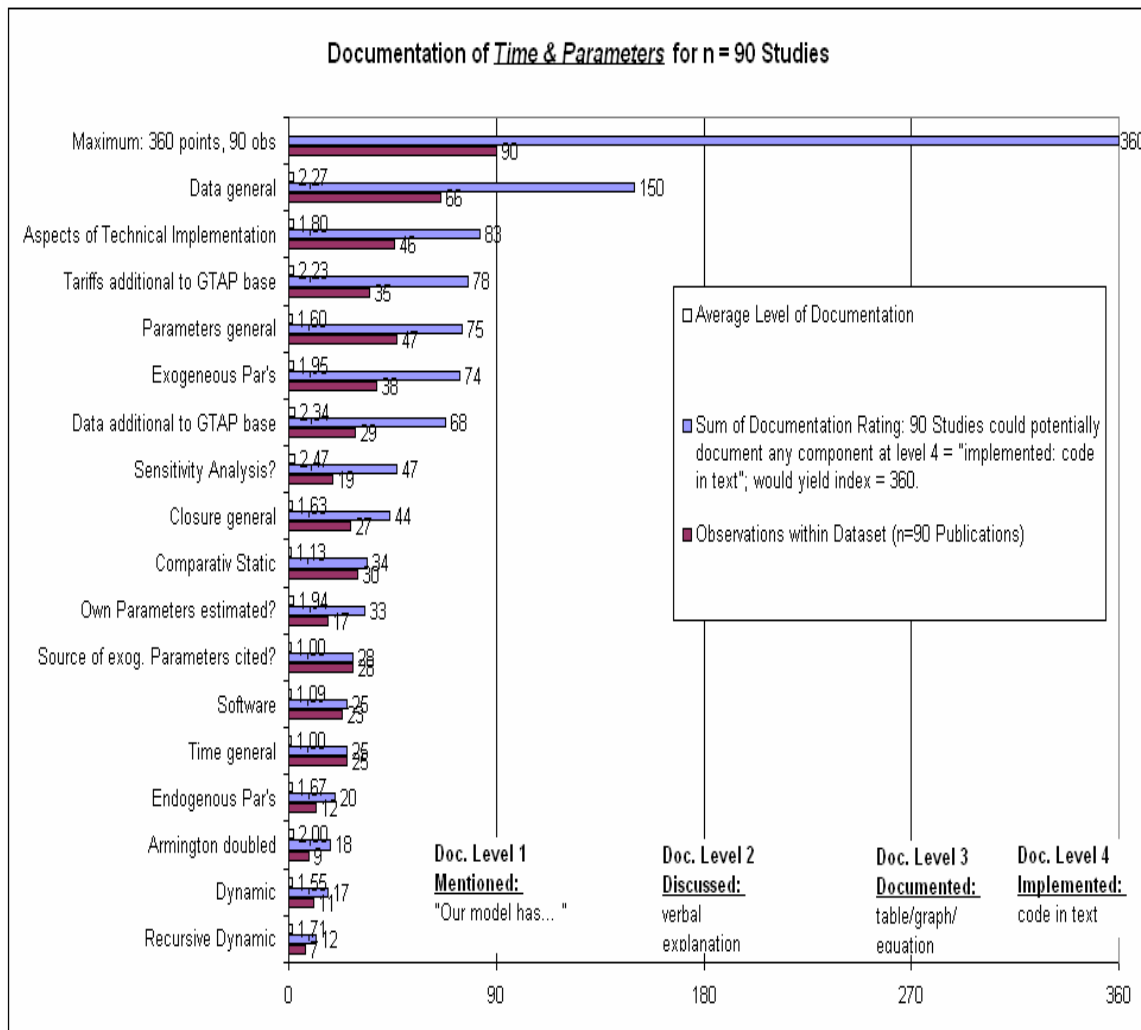


Table 1: Linear model of simulated welfare changes (EV in million US\$); OLS with robust standard errors.

Coefficients:	Value	Std. Error	t	Pr(> t)
Intercept	-11687320	375921.2	-3.109	0.00189
Model = BDS	181319	82287.1	2.203	0.02764
Model = GTAP	72348	43011.2	1.682	0.09266
Model = HRT	78952	33188.5	2.379	0.01743
Model = OWN	222061	79899.9	2.779	0.00548
Model = SMART(GTAP)	29284	26589.9	1.101	0.27085
Model = USMP	278374	96377.0	2.888	0.00390
Inc>Returns to Scale? (Y/N)	-73888	24505.4	-3.015	0.00259
CES Function (Production)	126932	37822.1	3.356	0.00080
Leontief (Production)	-49896	13987.2	-3.567	0.00036
New Growth Features?	-156083	43083.0	-3.623	0.00029
CET Function (Production)	-188241	56552.6	-3.329	0.00088
CDE Function (consumption)	83984	23640.2	3.553	0.00038
Regional Household?	-83943	20155.4	-4.165	0.00003
Sluggish Factors?	-47316	13959.5	-3.390	0.00070
Imperfect Competition?	-237612	88069.2	-2.698	0.00701
Armington?	-158558	48366.3	-3.278	0.00105
Dynamic?	55341	24217.3	2.285	0.02237
Projection to Year	5726	1836.7	3.118	0.00184
No of Regions	-1837	607.1	-3.027	0.00249
No of Sectors	6381	2469.5	2.584	0.00981
Additional tariffs incl.?	-47773	11967.4	-3.992	0.00006
Other data included?	231079	71823.5	3.217	0.00130
Tariffs shocked only?	17879	6885.0	2.597	0.00945
Swiss formular experiment?	-13322	5075.6	-2.625	0.00871
Armington doubled ?	117629	39470.3	2.980	0.00290
Own Elast. Estimates incl?	233823	77818.5	3.005	0.00268
Other technic. Features ?	-91643	33731.1	-2.717	0.00663
Med. Term Policy optimistic?	122085	33360.2	3.660	0.00025
Global Shock (Y/N?)	63880	15806.4	4.041	0.00005
Residual standard error: 48600 on 2797 degrees of freedom				
Adjusted R-Squared: 0.3026				

Table 2: Descriptive Statistics of the Variables within the Regression Model.

Variable	Min	Mean	Max	SD
Welfare Change (EV, mill US\$)	-40672	6941.4534	2120510	58201.492
Projection to Year	1997	2001.3121	2013	5.652
No of Regions	8	22.5680	160	21.041
No of Sectors	0	17.3398	36	5.271
GTAP (Y/N?)	0	0.5515	1	0.497
Inc. Returns to Scale? (Y/N)	0	0.3134	1	0.464
CES Function (Production)	0	0.4577	1	0.498
Leontief (Production)	0	0.3629	1	0.481
New Growth Features?	0	0.2777	1	0.448
CET Function (Production)	0	0.0085	1	0.092
CDE Function (consumption)	0	0.2791	1	0.449
Regional Household	0	0.3477	1	0.476
Sluggish Factors?	0	0.3247	1	0.468
Imperfect Competition?	0	0.2193	1	0.414
Armington?	0	0.6548	1	0.476
Dynamic?	0	0.1012	1	0.302
Additional tariffs included?	0	0.5051	1	0.500
Other data included?	0	0.4885	1	0.500
Tariffs shocked only?	0	0.8426	1	0.364
Swiss formular experiment?	0	0.0407	1	0.198
Armington doubled ?	0	0.0732	1	0.261
Own elast. estimates incl?	0	0.2957	1	0.456
Other technic. features ?	0	0.1946	1	0.396
Med. term policy optimistic?	0	0.4340	1	0.496
Global shock (Y/N?)	0	0.6342	1	0.482