Replies to Referee's Comments on `gets` for *Stata Journal*:

The referee raises a number of valid and important points which I have addressed in the way described in the following paragraphs. There are four substantive points (numbered 1 to 4 in the referee's report). Action has been taken on all four of these to aim to address the shortcomings. The response to each of the points is described in order below. The referee has also raised a larger number of minor points. Each of these is indeed a valid concern, and has been corrected in line with the referee's suggestion. No further comments will be made on these points, however one clarification will be made in the final paragraph of this document.

The referee points out that the article appears to be directed to an econometrics or economics audience, and lacks required details for the broader audience that comprises *SJ* readers. This is a valid point, and a considerable shortcoming with the original submission. The revised article is (I hope) much more suited to *SJ*'s broad readership. The 'Introduction' section has been largely rewritten under the assumption that readers have no background knowledge of model selection nor general to specific modelling in general. I do maintain the assumption that readers may have an interest determining important predictors of a dependent variable using a large number of independent variables, and hence could find use in a GETS modelling process. This is most manifest in the assumption that readers will be interested in some clas of regression analysis. The article now provides a breakdown of key points in the GETS literature, and a (hopefully agnostic) discussion of some of the arguments for and against this type of modelling, so that readers unfamiliar with the history of this topic are aware of both its uses and limitations. As a result of the rewritten introduction, some technical definitions previously located in the introduction have been moved to the first paragraphs of the second section. Finally, as requested by the referee, discussion is provided regarding the use of `gets` in a wider range of circumstances. Some examples and references of use in the non-economics literature are provided in the introduction, and an additional paragraph is provided in conclusion which discusses the broad potential of GETS in Stata.

A second point is made by the referee regarding a seeming lack of context in the original article. Particularly, the article introduced the idea of the 'LSE approach' to econometrics, and key authors in this literature without the appropriate background. This is a point well taken, and the updated introduction

has removed this ambiguity and added additional references where necessary. Additional referals to this throughout the article (such as that flagged by the reviewer on page 10) have also been removed or updated for a wider audience. The particular phrase highlighted by the reviewer: 'perhaps Hendry' was not intended as a sly comment, but rather aimed to suggest that Hendry was a key author in this literature, and could perhaps be considered as the principal proponent of this methodology. However, I acknowledge that this unnecessarily muddies the water, so have removed it entirely.

It is pointed out that it is not clear that the functionality of the `gets` program is entirely based around Stata's linear modelling commands `regress` or `xtreg`. This is indeed an important point to make clear to readers, and in the revised article I have aimed to make this much clearer. Of particular note is the last paragraph on page 2 which is dedicated to this point, and a line in the abstract. Similar changes have been made to the help file, and this will sent to the SSC (along with some other minor changes pointed out by the referee) as an updated version in the coming weeks.

Finally, the referee makes the point that there is important discussion (such as that in Harrell's textbook) with regards to penalties that should be imposed on data mining methods such as GETS, and the possibility that key statistics (such as confidence intervals) may be underestimated. The revised paper discusses the important drawbacks and concerns involved in GETS modelling (see for example paragraphs 3, 4 and 5 of the article), with particular attention to these points from Harrell, and a number of other objections from Peter Kennedy. Whilst I flag these difficulties as important points for readers to be aware of, I do not go into great detail regarding solutions. I think that an article such as this is not the place for very detailed discussions of the econometric background, however I point the interested reader to a number of review articles and important papers which make points on both sides of the argument.

As discussed in the opening paragraph, this was a very careful review and the referee has made a number of other more minor suggestions. These are appreciated, and have all been adopted. The exception to this is with the referee's query regarding multiplication signs. I have kept these multiplication signs in the equation in the paper. This was simply a stylistic choice, as the combination of scalars directly next to italicised variable names in the equation looks somewhat awkward. However, I am not particularly attached to this decision, so if either referee or editor feel that these symbols should be removed, I am more than happy to do so.