Interesting paper. Interesting examples. Wide-ranging review of matching literature that will be of use to experts and novices alike. On the other hand, some methodology presented here as summary of advice from the literature is the invention of the authors, and not all of it strikes me as sound advice. I would much like to see this paper published in Political Analysis; I would also much like it if first certain methodological proscriptions were rethought and somewhat revised.

1. A complaint about the treatment of the example in Sec 7.1.
With one-to-one matching that throws out a good fraction of the controls, estimates can vary a good deal with small changes to the implementation of the match -- precise specification of the score, decision as to which variables to match exactly on, and even (in the case of greedy 1-1 matching, which it sounds like is what the authors are doing) the ordering of the lists of treatment and control subjects. This is discussed somewhere in Smith 1997 *Sociological Methodology.* So in order to make a fair comparison to Carpenter's Model 1, Figure 1 (p. 34) should try out a couple of these variations and produce a couple of solid-line curves, to reflect differences in the match one could end up with following their basic routine, as well as in the data model that's fit to the matched data. I would suggest that this display and accompanying discussion be updated to reflect uncertainty in the matching, in this way or in some other way of the authors' devising.

2. Second complaint about example of section 7.1: grounds for choosing between pair matching and multiple controls- or full-matching are distinctly misrepresented. Authors' reason for choosing one-to-one matching over one-to-many or full matching (p.32) is a poor one. They imply that these methods would not address issues of common support. In fact, either of these methods would better address the problem of common support than pair matching can, as they would permit the authors to use every observation within the region of overlap on the estimated propensity score, as opposed to that just those that could be placed into pairs by the matching algorithm. To address concerns about common support from within this framework, or actually from within either the framework of matching with multiple controls or pair matching, the authors could impose a propensity caliper.   The next step would be to get rid of the observations without a counterpart within caliper distance -- likely to be many fewer than the 102 observations that the authors' pair matching discards, at least if they select a reasonable value for the caliper.   Then one could one-to-k match, varying number of controls match, full match (with or without restrictions), or whatever. With the latter procedures one could match every last subject that had some counterpart in the other group within caliper distance on the estimated propensity. That is almost certainly not what the authors achieve with pair matching. Most every methodological paper on flexible matching that I know of would suggest that there are substantial gains to be had from introducing flexibility; it lets you match more people, and match them better.

If the authors elect to retain their pair matching, then I would suggest that they provide a rationale for doing so that is consistent with the methodological literature on the choice between pair matching and more flexible forms of matching.

Having made my two complaints, I think it's a well-chosen, interesting example, one that will enrich the literature. In fact, I'd suggest that authors make publicly available these data and their analysis of it, if possible. That would greatly increase the potential impact of the paper as well.

3. Throwing away observations. But let me expand upon this last criticism. It's one thing to reject an observation because it demonstrably does not belong; another to reject it because one was unable to shoehorn it into favored methodology -- as one does, in effect, when one sets aside an observation not because one could find no well-matched counterpart for it, but because a particular matching technique, in this case greedy pair matching, didn't happen to place it in a matched set. Fit the shoe to the foot, I say, not the other way around. To elaborate this distinction with an analogy, here is a Stat 101 example of rejecting an observation because it demonstrably does not belong: in a simple regression analysis, closer inspection of an outlier reveals that its data were incorrectly recorded. Corresponding example from Propensity Score Matching: estimated propensity score of a control observation differs greatly from that of any treatment observation. Example from Stat 101 of rejecting an observation because it was methodologically inconvenient: closer inspection of an outlier reveals that it's properly recorded and it absolutely belongs with the data set, but boy does it ruin the convenient appearance of a linear trend and inflate the standard errors. Example from Propensity Score Matching: I wanted to do a 1-1 propensity score match and my treatment group was twice the size of my control group; therefore I got rid of half my treatment group subjects. This is what's done in section 7.2 of this paper. Is it the case that every treatment subject who was rejected lacked a viable control anywhere in the data set? Perhaps, but almost surely not.

For many years, Rubin and coauthors have promoted something *similar* to what the current manuscript advises, namely 1-1 matching that matches only a small fraction of potential control subjects followed by parametric adjustment for the comparison of the treatment group to matched controls. See e.g. Rosenbaum and Rubin's 1985 *The American Statistician* paper, or Rubin 1991 *Biometrics*. Similar, but importantly different. There is a key difference between the setting Rubin envisions and that treated by current authors, namely, in Rubin's setup matching is performed when some needed piece of data, e.g. outcome or exposure status, has yet to be collected for the control group. In other words, Rubin *does not* recommend throwing away observations that are ready to be used; rather his 1-1 matching is in the service of deciding whom to focus data collection efforts on. I believe the 1991 *Biometrics* paper specifically disavows 1-1 matching for such a setup, proposing subclassification instead. For a noteworthy paper inveighing against the rejection of treatment group subjects in particular, see Rosenbaum and Rubin 1986, *Biometrics*.

Methods that often avoid unnecessarily rejecting observations include propensity-score subclassification, a la Rosenbaum and Rubin 1984 *JASA* or Imai and Van Dyk 2004 *JASA*, and matching with a variable number of controls, Ming and Rosenbaum 2000 and Bergstrahl and Kosanke 2003. Methods guaranteed to reject only those observations that the statisticians flags as in need of rejecting, perhaps because they have no counterpart within caliper distance, are full matching, a la Rosenbaum 1991 *JRSS-B*, or full matching with restrictions, as in Hansen 2004 *JASA*.

4. Approaching the simplicity of pair-matched analysis without waste of observations. To use these alternatives to pair matching in practice, one needs a method of analysis that is adaptable to the matching/subclassification structures that result from them. The authors have addressed this already for the case of subclassification, on p. 29. In the case of full matching or matching with multiple controls, the clear analogue of authors' recommendations for 1-1 matched data is to form a subclass consisting of all subjects placed into 1-1 matched sets, another for all subjects placed into 2-1 matched sets, another from subjects in 1-2 matched sets, another for 1-3 matched sets, and so forth: then one fit parametric models separately in each subclass and take weighted averages of the results, as authors suggest on p.29. (Incidentally, wouldn't a reference to Imai and Van Dyk 2004 be useful at that point in the discussion? Don't they try out this analytic method with subclassification, with nice results from an accompanying simulation study?) In the example of section 7.2 I would guess that this would generate two to four subclasses: certainly a subclass for subjects placed by the matching routine into 1-1 pairs and a subclass for subjects placed into 2-1 matched triples; perhaps also a subclass of 1-2 matched triples, and/or a subclass of 3-1 matched quadruples. Accommodating these two to four subclasses within a basically parametric regression framework ought to work pretty smoothly. Another alternative many analysts would find painless would be to add a random effect for each matched set, a la Smith 1997 *Sociological Methodology*.

5. A gripe with authors' discussion of "the propensity score tautology" [p.23]
Reference to "consistency" is misplaced. Rosenbaum and Rubin 1983 *Biometrika* do establish of one particular "nonparametric" estimate of the propensity score, a sort of empirical histogram estimate, that in the limit as *n* goes to infinity it is a balancing score. There's nothing there like the result authors seem to assert, that for any old propensity estimate, parametric, semiparametric, or nonparametric, if it is consistent then it brings about balance. I don't know of any theoretical result, asymptotic or otherwise, that would entail that. If authors know of such a theorem, I would suggest that they provide a reference and explain the point a bit more.

Second, there's no hint in the literature that a score's having the balancing property would entail its consistency as an estimate of the propensity score, in large or in small samples. Quite the opposite; R & R 1983 clearly establish that propensity score is just one of countless balancing scores.

This should be carefully rephrased for accuracy. Basic idea seems right to me; with propensity scores you basically click your heels and hope to go to Kansas. Who knows whether there's really a wizard who's listening; point is that when you open your eyes again, you can readily check whether you've made it to Kansas or not, and surprisingly enough you often do.

6.The fallacy of a balance test fallacy. In my view, the paper is all wet when it says testing for balance is either impossible or misguided [pp.25-26]. I disagree, too, with the claim that "balance tests do not provide levels below which balance can be ignored"; but I'm going to save my breath for another day. If, as I suspect to be the case, there's overlap between authors of the present paper and the paper cited in the discussion, then this bit of tendentiousness is relatively innocuous, because readers will understand the authors to be restating an argument they make elsewhere rather than summarizing facts that are widely agreed to. I do think that *some* citation of other perspectives could improve the paper, but only if it were a credible other perspective. Unfortunately I don't have handy a citation that speaks precisely to this issue.