TV, Female Empowerment and Fertility Decline in Rural India: Response to Iversen and Palmer-Jones

Robert Jensen
Wharton School
University of Pennsylvania

Emily Oster
Booth School of Business
University of Chicago

Abstract: Iversen and Palmer-Jones (I-PJ) have undertaken a replication analysis of our 2007 QJE paper (JO) on the effects of cable television on women's status in rural India. The authors were able to replicate our analysis. They identified a small coding error which made no substantive difference to our results. They then introduce a series of adjustments, including constructing indexes differently, excluding some variables and exploring heterogeneity of the effects (some of which we had done ourselves in our 2007 NBER Working Paper). We recognize the scientific merit of replication and fully support such efforts. We also appreciate the willingness of the authors to take on this (often contentious, and thankless) process, as well as the considerable amount of effort the authors have put into their study. However, we disagree with the content and conclusions of their replication, and emphasize that their results do not change the conclusion that cable television appears to have a large and statistically significant effect on fertility, girls' education, women's autonomy, attitudes towards domestic violence, and (negatively) son preference. In fact, some of the claims raised by the authors arise from errors in their analysis. Finally, we make a few points about the process itself that we hope can improve future replication activities.

1. Introduction

We appreciate the opportunity to respond to the Iversen and Palmer-Jones (hereafter, I-PJ) critique of our paper on the effects of cable television on women's status in rural India (Jensen-Oster 2007, hereafter JO). We would like to start out by saying that we completely agree with the aims and value of replication. There is no need for us to argue here the virtues of replication, as they are well documented in 3ie program documents and elsewhere. We therefore worked to be as supportive as possible of this effort, including providing the data within a day of receiving the request, and providing timely responses to all queries.

Replication is by its very nature a process that is unlikely to win anyone a popularity contest, so the authors should be lauded for their willingness to undertake this effort. We also appreciate the obviously considerable amount of work they have put into this replication. As we will discuss more below, this replication did locate a small error in our code. While it made no difference to our central conclusions, we were grateful to have it bought to our attention and we immediately posted a corrected table and figure on our websites (this back in summer 2012).

We wish to emphasize that we disagree completely with the conclusions offered by the authors that our results are “weakened” and “do not provide strong evidence” or "are not robust." To be sure, we are grateful that the authors pointed out a coding error to us, which we subsequently corrected, and which, importantly, has no bearing on the conclusions at all. We are also grateful that they have made us realize that another variable was mislabeled (more on this below); again, this has no bearing on the conclusion at all. However, in every other way, and despite the way the authors have stated their findings, the conclusions of our paper remain clear and unchanged. After controlling for village fixed effects and time trends, there is compelling evidence that villages that added cable TV experienced improvements in measures of women's autonomy, attitudes towards domestic violence, schooling, fertility and reported son preference. These results can be seen in graphs with sharp changes and no pre-trends, and are robust to a variety of checks and tests.

We begin below by addressing what we see as the major claims of the authors. In section 2, we discuss concerns IP-J raise about the robustness of two key indexes used in our analysis of the SARI data. In section 3, we address concerns they raise about our use of the DISE data. Section 4 analyzes their discussion of mechanisms and theory of change, which entails examining how the results vary by education and television owning and/or watching. Section 5 addresses some of the smaller issues raised by IP-J. Finally, we make some points about the replication process itself, which we hope might assist in making future replications both more efficient and fruitful.

2. SARI Data: Index Construction and Robustness

The primary issue that IP-J raise with our SARI results is the construction of two index variables. We should note first that these issues relate to only two of our five outcome variables; our measures of fertility, education and son preference are based on single variables.

In our analysis of the effects of cable on women's autonomy and women's reports of the "acceptability" of a husband beating his wife, we combine answers to many different questions into a single index for each of these outcomes. We do so by averaging. IP-J argue that more might be learned about the sensitivity and interpretation of the results by disaggregating. They undertake this in their Table 2. They first show the results for each index component separately, then for various subsets, and then they explore other methods of combining these variables other than averaging.

We agree that looking at the various index components separately is interesting. In fact, as IP-J note, we ourselves did this in an earlier version of the paper, available online as an NBER Working Paper (Jensen and Oster 2007). Editors and referees didn't find this analysis particularly illuminating so it did not make it into the final version.

The more significant complaints by IP-J surround the sensitivity of the index construction.

Autonomy Index

In the case of autonomy, IP-J argue, first, that we should not include both the variable about needing permission to visit family and friends and the variable about making decisions about visiting family, on the grounds that they seem the same. They bolster this claim with evidence showing a high correlation between the two (pg. 9). In fact, all of these variables are highly correlated and these two are not especially closely linked. Of the 15 pairwise correlations between the six index components, this correlation ranks 7th. It therefore doesn't strike us as especially disciplined to remove these particular variables from the index. But perhaps just as importantly, IP-J note that, even with either of these variables removed, the impact of cable TV introduction on the index is still significant at the 1% level. So the conclusions of our analysis are unchanged.

They then show what happens when, instead of averaging, they use a PCA or MCA analysis to combine the variables. They note that in the MCA analysis, when they exclude one of the two "objectionable" variables the impact of cable TV is no longer significant. They say that this indicates that "the result of the JO specifications are sensitive to the way, in particular, the tolerance index was constructed."
On its face, this claim seems to us a vast overstatement. IP-J show 9 ways to construct this index (MCA, PCA and averaging as ways to combine the questions, and then for each of these three, either using all the questions, or dropping one of the two objectionable variables), and make the statement about sensitivity based on the fact that one of them is not significant. However, it turns out that even the claim about the lack of significance for this one permutation is also incorrect, and is the result of the authors making one additional adjustment to the data.

IP-J state, on page 12, that they construct their indices "Using the same component questions in index construction as JO." However, when we reviewed the Stata code they provided (because we could not replicate the results), we found that they had changed the binary variables we use for participating in decisions into ones which took on values of 0, 1 or 2. When they averaged the variables they used our binary definitions, but when they did the PCA ad MCA indices, they used their new variables. One could certainly debate which is the appropriate one to use, but the text is clearly misleading, and at least involves one additional change to the data.

The table below replicates the relevant portion of the IP-J Table 2 (in the first column) and then shows the corrected results using the actual index components from our paper.

	IP-J Coefficients (Table 2)	Corrected Coefficients
Autonomy, PCA (6 components)	0.131***(.038)	0.168***(.040)
Autonomy, PCA (exclude (iii))	0.180***(.041)	0.195*** (.041)
Autonomy, PCA (exclude (iv))	0.086**(.041)	0.126***(.042)
Autonomy, MCA (6 components)	-0.054***(.026)	-0.117***(.028)
Autonomy, MCA (exclude (iii))	-0.094***(.028)	-0.105***(.022)
Autonomy, MCA (exclude (iv))	-0.026 (.031)	-0.097***(.032)

Thus, when done correctly it doesn't appear that this result is sensitive to index construction, even to the extent that IP-J claim. Even taking the possibility that one might consider using the pure numeric values of the questions rather than converting them to binary indicators, we are now left with the summary result that of 18 possible permutations in constructing the index (a. averaging the variables vs. MCA vs. PCA, b. binary vs. averaging, c. including all 6 questions vs. excluding question iii vs. excluding question iv), just one of them is no longer statistically significant. Counter to the claims of lack of robustness, this seems to suggest to us that in fact the results are extremely robust to alternative constructions. We imagine that for almost any variable one might use in any empirical analysis, there exist some combination of changes along various dimensions that would yield a result that is no longer statistically significant. Calling this a lack of robustness seems like a very strict standard to hold any study to.

Spousal Beating Index

IP-J show that if you generate an adjusted tolerance for spousal violence measure which excludes the component of the beating index which is most significant in individual regressions, the impact of cable TV introduction on that adjusted measure is significant only at the 10% level. We agree that this is true – there are no issues with their code – but believe there is no reason to drop this variable from the index.

The justification for doing so is that the value of reported acceptability of beating for this variable is much higher in one of our states than in the NFHS2, which asked a similar question. However, this is due to a labeling error on the variable in our data set. We are very grateful to the authors for raising the discrepancy and thereby making us aware of the error. In fact, the question in our survey is not the same as that in the NFHS, so there is no reason to exclude it based on "external validity" concerns.[1]

[1] We originally tested a question, based on the NFHS, asking whether it was acceptable for a husband to beat his wife if her natal family does not give expected jewelry, money or other things. However, after piloting our survey but shortly before taking it to the field, there was a tragic case involving an alleged dowry murder in Tamil Nadu. Because this case was receiving considerable media attention, we were worried that the salience of this issue might influence the responses to this question (or raise concern that our enumerators were part of a police effort), so we changed the question just prior to final printing. The label in the data set that we provided to PJ-I, and indeed the one we ourselves had for our analysis, is incorrect. The question was changed to whether it was acceptable for a husband to beat his wife if she took money that was meant to be used for other things in the household and misused it or used it for herself. Though even this is still much higher in Tamil Nadu than in our SARI states, the same is generally true of all reported spouse beating levels.

We also note that even were the questions the same, there would not necessarily be a clear rationale for excluding it on the basis of differences in responses in the two surveys alone. Responses to similar questions could arise for many reasons, including being conducted in different years (as is the case for our data and the NFHS), current events, differences in the ordering of questions, sample composition or just randomness or sampling variability. And for example, in the NFHS data the authors show to raise external validity concerns about the SARI data, a similarly large fluctuation arises even within the two NFHS surveys. The bottom two panels of IP-J's Figure 2 show data from the 1998/9 NFHS2 and the 2005/6 NFHS3. Note in particular the tallest blue bar in the bottom panel, representing the acceptability of violence towards a wife if the husband suspects she has been unfaithful in Haryana, measured in the NFHS3. Note that in the NFHS2 survey, the mean was just over 20 percent, compared to about 70 percent in NFHS3. There is also an increase in Goa as well, but it is much smaller. The two surveys are nearly identical and intended to be completely comparable, so again, a single outlier variable should not necessarily be taken to raise external validity concerns; certainly not enough to warrant excluding a variable from the analysis.

3. DISE Data: Sensitivity of Results

The second issue we take up is the sensitivity of the DISE results to errors or additional controls. Although the paper is a bit opaque, we think it is useful to outline and address the three issues raised separately.

Coding Error

There were two small coding error in our paper, which were brought to our attention through this process. As noted above, we very much appreciated this. In July 2012 we posted a Corrigendum in our websites, which featured an updated Figure 7 and Table 6. There were in fact two errors:

1. We had included 5 years olds in the under 10 group, but not the overall under 14 group in years 2005-2007 . This resulted in, in some cases, us calculating a negative number of 11 to 14 year olds. Since we took the log, these cases were then dropped from the analysis.
2. In calculating the total fixed cohort in 2007-2008 we had included 10-year-olds and 12-year-olds rather than 11 and 12 year olds.

In the end, this resulted in small changes to Figure 7 and Columns 1, 2, 5 and 6 of Table 6. The quantitative changes were small. One figure (Panel B, Column 1) was significant at the 10% level and is now not significant. Two others (Panel C, columns 5 and 6) are now larger in magnitude and significant. Our overall conclusions were unchanged by this. The other results with the DISE data (Columns 3 and 4 of Table 6 and Appendix tables in which we analyze data by class rather than age) are unaffected.

Disagreement on Figure 7

IJ-P devote a long discussion to our disagreement on the construction of Figure 7. The small coding errors noted above did affect Figure 7 but only in an extremely minor way. The bigger issue – the larger reason why our figures differ from theirs – has to do with the construction of a balanced panel.

The DISE data mostly covers primary schools. In some cases these stop at grade 5. Given the length of the panel we are considering in Figure 7, in a number of cases the enrollment went from a large positive number to zero at a time when it appeared that students may have aged out of the primary school. Our construction of the figure does not include these schools, since we were concerned that "0" enrollment did not reflect no students in the school but, rather, the fact that the students that were there graduated. IP-J do include these schools, which is what pulls down their enrollment in later years.

In the end, this issue seems to have taken outsize importance in the discussion. In the primary empirical results – in our Table 6 – we include all schools, not only those in this limited sample. We are in much better agreement with IP-J on the table than on the figure, and this seems the more central test.

Building Quality Controls

The final issue that IP-J bring up is that – in a scientific replication approach – they include controls for building quality (and for English-language schools in the village) and find this eliminates our results. We could not fully replicate the results in their Tables 8 and 9 – running their code as provided yielded slightly different results for all coefficients – but we were able to replicate the general fact. It is perhaps important to note that it is not only a control for building quality which matters, but this control interacted with year. Controlling for level of building quality (which does vary across years) does not impact our conclusions.

Our primary concern is that the building quality controls used include, among other things, measures which effectively capture the *size* of the school. This includes number of blackboards, books in the library, etc.

By putting these on the right hand side when enrollment in on the left, they effectively change the analysis. The outcome is now effectively measuring density of students per classroom rather than number of students. We certainly would not – and do not – argue that cable TV impacts classroom density. In fact, we considered this issue in our initial paper but ultimately concluded the best we could do was to control for electricity – an element of building quality which is not associated with size – which we do in our primary results.

Given this size issue, we find this analysis to be misleading and the conclusions drawn from it – namely, that "the DISE data analyzed in these ways do not provide strong evidence in support for the assertion that Cable TV increase enrollment." -- to be flawed.

4. Mechanisms

A large portion of the IP-J paper is focused on trying to better understand the mechanisms by which the effect we see may be occurring. This is done, in part, by looking at interactions between television access and demographics. As IP-J note, we had done some of this in an earlier working paper version. They expand on this analysis considerably.

We think this is interesting and useful. IP-J imply several times that we should have included this in our paper, since we had it in the working paper, and we will say in response only that in the publication process papers naturally become shorter and more concise. The referees and editors did not feel that these results were especially informative, and we therefore dropped them.[2]

We have limited comments on this other than to say that we think it's worth taking some of the other results, which interact access to TV with TV watching behavior, with a grain of salt, since the watching behavior is surely endogenous.

5. Smaller Issues

A. The authors express concern that our analysis considers attitudes towards domestic violence, rather than direct measures of the experience of domestic violence. In particular, they state that:

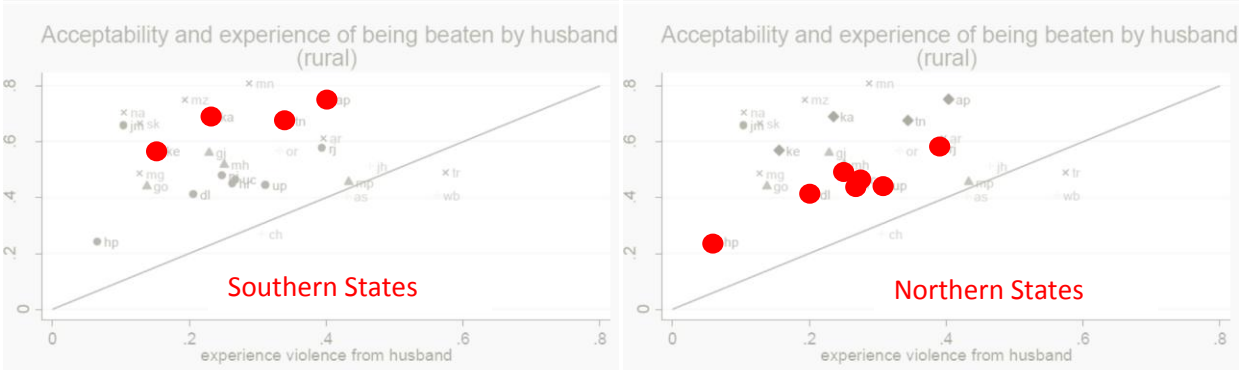
a. More direct measures are recommended, unless attitudes are correlated with actual.
b. In the NFHS, "the two are not always well correlated" that is not "an insignificant matter" and there is "an apparent if limited association between experience and acceptability other than in the east."

First, as the authors note, we acknowledged this point ourselves in both our working paper and the published version. Attitudes are certainly not the same as underlying behaviors. Unfortunately, we have no direct measure of the experience of domestic violence, and recognize our measure as limited. However, as we argue in the paper, there may be some value in the measure we have: "...even if cable only changes what is reported, it still may represent progress: changing the perceived "correct" attitude seems like a necessary, if not sufficient, step toward changing outcomes."

We deliberately did not ask about direct experience of spousal violence in our survey. This was primarily out of concern over the respondents, given the potential harm involved in asking someone to divulge and talk about or re-live a traumatic experience, or potentially exposing them to the risk of further harm. But we also note that it could affect data quality. For sensitive subjects, people may be more likely to underreport these behaviors; this could even have a chilling effect on the quality of the rest of the survey. One might even argue that asking about attitudes is likely to yield more meaningful responses. But ultimately, we acknowledge that for such a sensitive topic, there is no good answer. Even official police statistics are not likely to be valuable, since domestic violence is surely underreported.

[2] One concern is that variables such as education may proxy for many other things (for example, wealth, family background, local attitudes towards women, how traditional an area is, other infrastructure, etc.) so the interpretation is unclear.

But setting this aside we disagree with their claim in point (b) above about their being a limited relationship between the attitudes towards domestic violence and the self-reported experience of domestic violence in the NFHS3 data the authors consider. The authors argue that their Figure 1 shows that this relationship, measured by averages at the state level, is weak, other than in eastern states. First, despite being a very small sample, there is to our eye a clear relationship in Figure 1. There are not many data points, but there certainly appears to be a relationship between the two overall, as well as in the northern states and southern states, where the SARI data were collected, which becomes clear when the two are isolated:



Just as important, state-level data do not seem to be the most appropriate way to explore this relationship. The table below estimates regressions at the individual level using the NFHS3, considering four samples: All States, All States (rural areas only), SARI States and SARI States (rural areas plus Delhi). In accordance with the graphs the authors use, to gauge whether the two are correlated, we simply regress whether a woman reports it is acceptable for a husband to beat his wife in any situation on whether she reports having been physically abused by her husband (the results continue to hold when controlling for age, education, religion and other household variables). In all four, the relationship between reporting having experienced domestic violence from a spouse and reporting at least one situation in which it is acceptable for a husband to beat his wife are highly economically and statistically significant. The t-statistics range from 7 to 20. Of course, the two are not perfectly correlated. However, the strong correlation exists. Similar results hold for having experienced "emotional violence" as well.

	All States	All States--Rural	SARI States	SARI States--Rural+Delhi
ok_beat_any	0.107***	0.085***	0.144***	0.151***
	(0.005)	(0.007)	(0.015)	(0.019)
Constant	0.298***	0.334***	0.367***	0.371***
	(0.004)	(0.005)	(0.011)	(0.014)
Observations	69,390	38,904	11,082	7,060

The dependent variable is having ever experienced any form of abuse by a spouse. Heteroskedasticity consistent standard errors reported. Regressions also make use of sampling weights to account for the fact that the domestic violence experience module was only administered to a subsample of women. Results without the weights are even stronger.

Again, we would not use these results to claim that we have a good measure of domestic violence. In fact, our measures are not intended to be measures of domestic violence itself. However, the lack of a

correlation between these attitudes and self-reported experience of domestic violence, given as a reason for greater concern about our results, is not borne out by the data.

B. Drawing on Basu and Koolwal (2004), the authors argue that some measures of autonomy should be more easily changed than others. They then analyze whether there are spillover effects of TV (effects for those who do not watch TV) for components of the autonomy variables that they classify as hard to change or easier to change. They conclude that: "Excepting the 'bad cook' variable from the tolerance index, the spillovers for the autonomy variables appear for two of the remaining three hard to change variables. This fuels concerns about potential and time-variant confounding factors, rather than the arrival of cable TV, as the underlying driver of social change." Again, we feel that the stated conclusion here is far too strong. We feel that the classification of some variables as hard vs. easy to change is subjective and arbitrary, even when drawing on the framework of one previous study. Certainly, they could be debated--but finding a change in variables that one considers to be hard to change hardly seems to merit the conclusion that something else must be causing the change other than TV.

6. Process and Timeline

We again would like to emphasize that we view replication as a valuable part of the scientific process, and we appreciate the considerable efforts of Iversen and Palmer-Jones. However, we wish to emphasize the need for timeliness, respect and communication when undertaking a replication of this type. This seems especially relevant since 3ie has commissioned a number of other replications with this as the model. We make some of these comments with respect to our specific experience, whereas with the others, they are suggestions that arise as we think through the potential issues more generally.

We were first asked for our data by IP-J in June of 2011. We sent them the first of our data (the SARI data) within a day, and had all of the data to them by a week later. We then exchanged a series of emails to clarify (a total of twenty or so) at the end of July 2011. IP-J concluded this exchange with telling us they would be in touch with any further questions when they sent a draft of the replication. We then heard nothing for about a year. In May of 2012 one of us sent an unrelated email to IP-J and asked about the replication in passing. They mentioned to us that they had located a small error in our code. It was at this point we fixed the error and generated an altered table and figure. They assured us that they would send us the draft replication to comment on before circulating.

However, in July 2012 we were made aware of the fact that the authors had already given a public seminar about the paper. When we subsequently explored the slides, they contained numerous errors (which have mostly now been fixed) which we could have identified had we been contacted about them, as it was suggested we would be. Although subsequently the authors agreed not to circulate the paper until it was completed, we nevertheless felt at least some of the damage had been done.

It was another full year until the replication process was completed and we were allowed to see the final paper and to comment. Some of this was due to delays clearly external to IP-J, such as a draft being sent to reviewers for comments (but again, we feel when a replication paper is being sent to reviewers, it is well past the time when the original authors should have been given a copy of the paper--even if only to save the time of the authors and referees, who might be alerted to mistakes). At this point we were given approximately 4 weeks to prepare a response. This is after the original authors had two full years. This required us to drop other projects. We had asked repeatedly over the last year to see the draft, as it

was sent out to reviewers, revised, and sent out again. We were never given access to it to begin preparing our response.

We view this as somewhat of a cautionary tale and would urge 3ie and other organizations undertaking replications to consider that the time and reputation of those whose papers are being replicated might be given similar respect to that of the replicators.

3ie has taken the laudable step of making funds available for authors of studies being replicated to facilitate the preparation of data sets to be provided to those undertaking the replication. In our case, the data were already prepared (and had previously been given to several other researchers), so no such funding was necessary (and this offer was only extended after the replication had begun). However, perhaps just as important to most researchers is the time cost of such efforts. Providing replicating authors with and making them adhere to strict deadlines, providing inputs and feedback timely, and keeping in communication on timeline and process are valuable as part of reducing the burden on the original authors. This could also greatly enhance the efficiency and effectiveness of the replication, without sacrificing independence, thoroughness or quality (in fact, likely improving it).

It should also be emphasized to replicating authors that the goal of a replication is not to overturn the results of a previous paper. Authors should go in with the mindset of making sure the results can be replicated as they appear in the original paper, are robust to plausible alternatives, search for sources of external validity where available, etc. The incentives of the replicators, particularly in terms of publication, are to "overturn" the original results, and could lead to overstatement of the magnitude of criticism. In part, 3ie has contributed to solving that problem by providing financial resources to the authors of replications, whereas in the past the only reward was to hope for publication of a comment, which again would be contingent on overturning the original result.

The distinction may seem subtle, but is important. Just as a researcher producing original research should not set out with the belief that their hypothesis is correct, nor should we start with the presumption that it is wrong. Both can lead to intentional or even unintentional or subconscious efforts to prove one's hypothesis, either by ignoring supportive results, or specification searching, cutting data along various dimensions and constructing alternative approaches until one finds a null result. Ultimately, if one searches long enough, a way to undermine any result can be found. Though replicating authors should be rigorous and thorough, there may be a fine line between plausible alternatives and data mining. Part of this problem could be addressed with the use of a "pre-analysis" replication plan, as 3ie now appears to require--in the present case, to our understanding, a replication plan was not prepared in advance. But in general, a repeated emphasis on the purpose and process of replication would greatly assist future replication efforts.

References

Basu, A.M., Koolwal, G.B. (2005). "Two concepts of female empowerment–Some leads from DHS data on women's status and reproductive health." Focus Gender–Collected Papers on Gender Using DHS Data. ORC Macro: Calverton, Md.

Jensen, Robert T. and Emily Oster (2007). "The Power of TV: Cable Television and Women's Status in India," National Bureau of Economic Research Working Paper No. 13305, Cambridge, MA.

- - and - - (2009). "The Power of TV: Cable Television and Women's Status in India," *Quarterly Journal of Economics*, 124(3), p. 1057-1094.

Iversen, Vegard and Richard Palmer-Jones (2013). "TV, female empowerment and fertility decline in rural India," 3ie Replication Working Paper Series.