A Comment On:
School Choice: An Experimental Study[*]

Caterina Calsamiglia   Guillaume Haeringer[†]   Flip Klijn[‡]

August 8, 2009

Abstract

We show that one of the main results in Chen and Sönmez (2006, 2008) does no longer hold when the number of recombinations is sufficiently increased to obtain reliable conclusions. No school choice mechanism is significantly superior in terms of efficiency.

JEL classification: C70, C13, C91.

Keywords: school choice, efficiency, recombinant estimator, robustness.

We consider the experimental study of Chen and Sönmez (2006,2008 —henceforth CS for short).[1] CS's experiment was intended to assess the relative performance of three school choice mechanisms: Boston (BOS), Gale-Shapley (GS), and Top Trading Cycles (TTC). Their experimental study complemented the mechanism design approach of Abdulkadiroğlu and Sönmez (2003) to study the assignment of children to public schools in the US. As such it played an important role to convince the Boston school district authorities to replace the previous mechanism (BOS) by one of the other mechanisms.[2] The choice between GS and TTC mainly depended on the relative weight that the authorities assigned to stability versus efficiency. Abdulkadiroğlu and Sönmez's (2003) (theoretical) results are very clear: GS is stable (but not Pareto-efficient) and TTC is Pareto-efficient (but not stable). However, CS's "perhaps most surprising result ... concerns the efficiency comparison of the three mechanisms, as [their] experimental results do not support theory" (CS, 2006, concluding discussion on page 229). In particular, they find that GS is significantly more efficient than TTC. In this note we show that CS's claim does no longer hold when the number of recombinations is sufficiently increased to obtain robust conclusions. More precisely, we will see that no school choice mechanism is significantly superior in terms of efficiency.

[†]Departament d'Economia i d'Història Econòmica, Universitat Autònoma de Barcelona, Spain; `caterina.calsamiglia@uab.es`, `guillaume.haeringer@uab.es`

[‡]Institute for Economic Analysis (CSIC), Campus UAB, 08193 Bellaterra (Barcelona), Spain; `flip.klijn@iae.csic.es`

[1]See Sönmez and Ünver (2008) for a survey on recent developments about school choice mechanisms.

[2]See Abdulkadiroğlu et al. (2005) for a first report on the recent redesign of the Boston Public School match.

CS considered two environments (one based on a designed preference profile, and the other based on a randomly generated preference profile) and thus obtained 6 treatments.[3] For each treatment they ran two sessions (i.e., $n = 2$), with $k = 36$ students in each session. CS employed a recombinant estimation technique with $r = 200$ recombinations to obtain a refined analysis of the relative efficiency of the mechanisms. Their statistical analysis was based on t-tests. We describe the recombinant technique as well as the statistical estimators in Section 1. Finally, in Section 2, we show that when the number of recombinations is sufficiently increased in order to obtain stable conclusions CS's result that GS outperforms TTC can no longer be sustained.

1 Recombinant Technique and Estimators

Recombinant techniques are a useful tool to analyze data obtained from laboratory experiments based on normal form games.[4] The idea behind recombinant techniques is that as long as one is interested in the analysis of the outcome of the game (i.e., payoffs, not the strategies) running the experiment a "few" times suffices to obtain more experimental data. More precisely, for the game in CS one can generate up to $n^k = 2^{36}$ "virtual" data sets by picking each of the k players' strategies from either of the n sessions.

To avoid the computationally impossible task to calculate the outcomes induced by all virtual data, CS employed the recombinant estimator proposed in Mullin and Reiley (2006), which requires running fewer recombinations. In the case of the experimental data of CS the procedure boils down to the following. One starts by picking the strategy of the first subject from the first session, and then choosing randomly the strategies of player 2 up to 36 from either of the two sessions. For this strategy profile the outcome of the game is computed. Next, one repeats the procedure by picking the strategy of the first subject from the second session, and so on, until one has done so for all subjects from both sessions. As a general guideline, Mullin and Reiley (2006, page 177) recommend to repeat the procedure at least $r^* = 100$ times for each of the $n \times k$ subjects. CS opted for $r = 200$ recombinations.

Given the virtual data sets, CS compared the estimated mean payoff in each of the treatments in order to evaluate the efficiency of the different mechanisms. To determine whether the differences are statistically significant CS used t-tests based on the following estimators. Consider any of the 6 treatments. For each of its $n \times k \times r = 2 \times 36 \times 200$ recombinations, let $Y(i, j, l)$ be the *mean* payoff of the l-th artificial session created by fixing player j from session i. The estimated mean payoff over all recombinations is given by

$$\hat{\mu} = \frac{1}{14400} \sum_{i=1}^{2} \sum_{j=1}^{36} \sum_{l=1}^{200} Y(i, j, l).$$

The estimated variance in payoffs is then given by

$$\sigma^2 = \frac{1}{14400} \sum_{i=1}^{2} \sum_{j=1}^{36} \sum_{l=1}^{200} [Y(i, j, l) - \hat{\mu}]^2.$$

---

[3]See Chen and Sönmez (2006) for further details.

[4]See for example Engelbrecht-Wiggans, List and Reiley (2006), Apesteguia, Dufwenberg and Selten (2007) or Dufwenberg, Gneezy, Goeree and Nagel (2007) for recent applications of such techniques.

To compute the covariance, CS split each of the 200 recombinations $(i, j, \cdot)$ in two sets of 100 recombinations, and compute the covariance across these two sets, i.e.,

$$\phi = \frac{1}{7200} \sum_{i=1}^{2} \sum_{j=1}^{36} \sum_{l=1}^{100} [Y(i,j,l) - \hat{\mu}] \times [Y(i,j,l+100) - \hat{\mu}].$$

The asymptotic variance can then be estimated using Eq. (6.5) of Mullin and Reiley (2006),[5]

$$var(\hat{\mu}) \approx \frac{\sigma^2}{36 \times 200 \times 2} + \frac{36\phi}{2}.$$

2 Statistical Tests, Robustness, and Discrepancies

CS's choice to generate 200 recombinations per subject-session follows Mullin and Reiley's (2006) suggestion to use at least 100 recombinations (per subject-session). Nevertheless, it turns out that 200 recombinations is not sufficient to obtain robust statistics in such a rich game as the one representing each treatment. The results we obtained when we carried out multiple series of 200 recombinations vary considerably from one series to another. For each of the 6 treatments the mean payoff $\hat{\mu}$ and its variance σ^2 and covariance ϕ do not depend very much on the number of recombinations. But the asymptotic variance, which puts a higher weight on the covariance as we increase the number of recombinations, decreases with the number of recombinations, thereby affecting the results of the tests.[6] To give an idea of this variation we generated 150 series of $r = 200$ (resp. 2000, 10000, and 100000) recombinations (per subject-session) for each of the 6 treatments. Thus, in each case we obtained $150 \times 150 = 22500$ hypothesis tests for the 6 pairs of treatments. If the percentage of acceptance is 0% or 100% then the associated conclusion may be considered robust since all 22500 hypothesis tests led to the same conclusion. Table 1 summarizes the proportions of acceptance rates.

x vs. y \ r	200	2000	10000	100000
GS_d vs. BOS_d	99.87%	100.00%	100.00%	100.00%
TTC_d vs. BOS_d	28.37%	0.20%	0.07%	0.00%
GS_d vs. TTC_d	33.67%	40.78%	44.36%	33.28%
BOS_r vs. GS_r	0.00%	0.00%	0.00%	0.00%
BOS_r vs. TTC_r	28.11%	25.49%	6.98%	0.06%
GS_r vs. TTC_r	17.52%	18.89%	7.11%	0.01%

Table 1: Acceptance rates of H_0: $\hat{\mu}_x > \hat{\mu}_y$ and H_1: $\hat{\mu}_x = \hat{\mu}_y$.

[5]Abrevaya (2008) provides evidence that Mullin and Reiley's (2006) variance estimation can be downward biased and provides a method to avoid this bias. Our findings in the next Section are based on Mullin and Reiley's (2006) asymptotic variance but the qualitative results are also true with Abrevaya's (2008) method. If the result is that the difference between TTC and GS is not statistically significant with a downward biased variance, the difference will be even less significant with a larger, less biased variance.

[6]A first problem we encountered is that in many instances the estimated covariance from a given recombination was negative. That implied that the estimated asymmetric variance was negative. This problem disappears when the number of recombinations is increased.

Figure 1 additionally depicts the distributions of the p-values for 4 relevant cases. We omitted the cases GS_d vs. BOS_d and BOS_r vs. GS_r since it is clear from Table 1 that the associated conclusions are already very robust for $r = 200$. Note that in the remaining 4 cases the distribution of the p-values has a high variance when the number of recombinations is small. When $r = 2000$, TTC_d vs. BOS_d also becomes robust. For $r = 100000$ all results are robust except for GS_d vs. TTC_d. However, when $r = 200000$ this latter result also becomes (almost) robust.

How should we rank the mechanisms in terms of efficiency? For the designed environment, CS's corrigendum on Result 6 concluded that $TTC_d \sim BOS_d$, $GS_d > BOS_d$, and $GS_d > TTC_d$.[7] However, the values in Table 1 and the distributions in Figure 1 strongly suggest that in fact GS_d does not outperform TTC_d, i.e., $GS_d \sim TTC_d$. For the random environment, CS's corrigendum on Result 6 concluded that $GS_r \sim BOS_r$, $GS_r \sim TTC_r$, and $BOS_r > TTC_r$. However, Table 1 and Figure 1 provide evidence that in fact $BOS_r \sim TTC_r$.

As we have pointed out, we can *no longer* conclude that GS is superior to TTC in the designed environment (which in contrast to the random environment was specifically constructed to mimic a realistic environment[8]). In other words, our findings do not provide support to part of CS's "perhaps most surprising result ... [which] ... concerns the efficiency comparison of the three mechanisms, as [their] experimental results do not support theory" (CS, 2006, concluding discussion on page 229).

---

[7]Following CS's notation, $x > y$ denotes that x has a higher per capita payoff than y at the 5% significance level or less, and $x \sim y$ denotes that x does not have a higher per capita payoff than y at the 5% significance level.
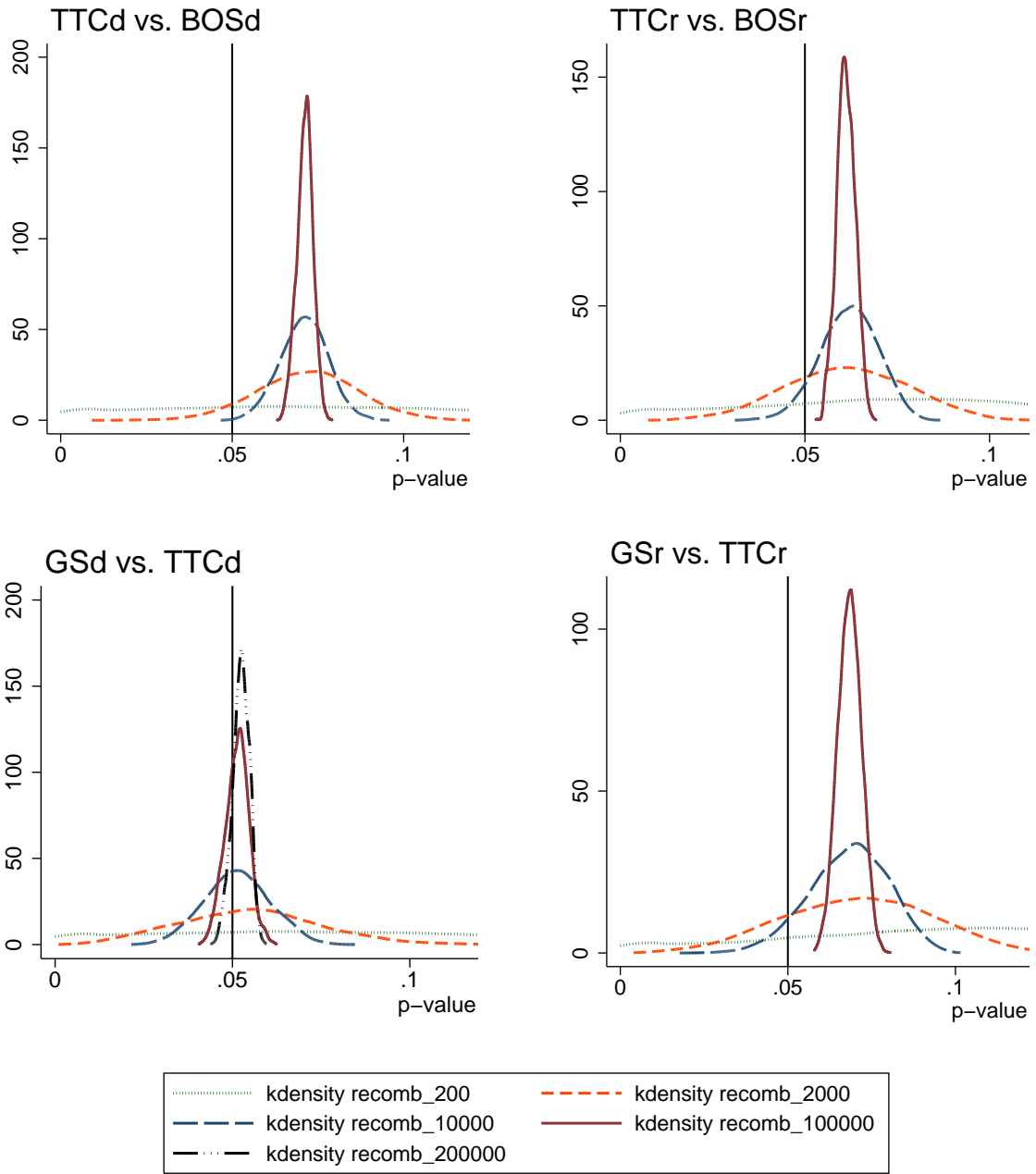
[8]See CS for details.

Figure 1: kernel densities of p-values

References

[1] Abdulkadiroğlu, A. Pathak, P.A. Roth, A.E. and Sönmez, T. (2005) "The Boston Public School Match," *American Economic Review, Papers and Proceedings*, **95**, 368-371.

[2] Abdulkadiroğlu, A. and Sönmez, T. (2003) "School Choice: A Mechanism Design Approach," *American Economic Review*, **93**, 729-747.

[3] Abrevaya, J. (2008) "On Recombinant Estimation for Experimental Data," *Experimental Economics*," **11**, 25-52.

[4] Apesteguia, J. Dufwenberg, M. and Selten, R. (2007) "Blowing the Whistle," *Economic Theory*," **31**(1), 143-166.

[5] Chen, Y. and Sönmez, T. (2006) "School Choice: An Experimental Study," *Journal of Economic Theory*, **127**, 202-231.

[6] Chen, Y. and Sönmez, T. (2008) "Corrigendum to School Choice: An Experimental Study," mimeo.

[7] Dufwenberg, M. Gneezy, U. Goeree, J.K. and Nagel, R. (2007) "Price Floors and Competition," *Economic Theory*, **33**(1), 211-224.

[8] Engelbrecht-Wiggans, R. List, J.A. and Reiley, D.A. (2006) "Demand in Multi-unit Auctions with Varying Number of Bidders: Theory and Evidence from a Field Experiment," *International Economic Review*, **47**(1), 203-231.

[9] Mullin, C.H. and Reiley, D.A. (2006) "Recombinant Estimation for Normal-Form Games with Applications to Auctions and Bargaining," *Games and Economic Behavior*, **54**, 159-182.

[10] Sönmez, T. and Ünver, U. (2008) "Matching, Allocation, and Exchange of Discrete Resources," in *Handbook of Social Economics*, Alberto Bisin, Jess Benhabib, and Matthew Jackson, (eds.), Elsevier.