Excessive similarity of coin size

The data in Wang et al (JDM 2012, 7(1),p.77-85)did not originate from random samples

The distributions of coin size estimates across experiments 1 and 2, for the conditions included in both experiments (shame & control), are excessively similar. The more extreme, shame:

Experiment 1 (n=25): 2,3,3,3,3,3,4,4,4,4,4,5,5,5,5,5,5,5,6,6,6,6,6,6,6,7
Experiment 2 (n=24): 2,3,3,3,3,3,4,4,4,4,4,5,5,5,5,5,5, 6,6,6,6,6,6,6,7

The rectangle highlights that if one were to delete a single observation from Experiment 1 one obtains the data for Experiment 2.

How likely is this to happen by chance?

We need an index for similarity of distributions across conditions. The obvious choice is the metric used for the chisquare test with contingency tables: $x^2=\sum \frac{(observed-expected)^2}{expected}$)

The x^2 value for the observed data is x^2 =.05653.
While x^2 is not distributed χ^2 because expected counts <5, we can use bootstrapping techniques to assess how unlikely low that x^2 is without relying on the $\chi^2(6)$ distribution.

We do the most conservative test possible by assuming both samples come from the exact same distribution (populations cannot be more similar than identical, and are almost certainly more dissimilar than identical, making the data seem more possible than they really are). We pool the data from both conditions, N=49, and draw with replacement one sample with n=25 and one n=24. We compute the x^2 value for those two bootstrapped samples, and repeat.

I conducted 200,000 simulations (takes ~ 9 minutes in R). An x^2 as low or lower than .05653 was obtained only 104 times. If we treat this as a test of the hypothesis that the data come from random samples, we reject that null, for the shame condition, with *p*-value=104/200000= .00052.

Similar calculations for the control condition, which has a χ^2=.34756, lead to rejecting random sampling with *p*-value=.0129. Note that both conditions involve different subjects, so they are two independent tests of data impossibility.

To get an overall sense of improbability of *both* conditions being so similar across experiments I added up the two χ^2 values (.05653+.34756)=.404, and assessed how many of the 200k simulations had the sum of both $\chi^2 \leq .404$. *Five* out of the two hundred thousand did, leading to an overall *p*=.000025

It is worth recalling that this is an incredibly conservative test, there is no reason to believe the two samples come from the same population (e.g., sample 1 is students, mean age 21, sample 2 is community members , mean age 31); .000025 is a vast overestimate of how probable the results are to arise from random samples