# Update Document April 2024 Re-randomization

# Courts of Tomorrow

**Experimental Design**

This report outlines the methodology of our randomization scheme for judges in Pakistan. The random assignment of judges were conducted in two distinct phases. Initially, in February 2024, we randomized 980 judges (who consented to be part of our research). These judges were then randomly assigned into two groups: the first group comprising 488 judges, was designated as the treatment group (Batch 1) and received access to the AI course and JudgeGPT subscription in February 2024; the second group, consisting of 492 judges, was to be randomly assigned to be control group (Batch 2) and would be treated in November 2024, at the time of our endline.

After the initial random assignment, the introduction of the password-protected JudgeGPT, which was designed specifically to prevent spillovers in the experiment, sparked significant interest among judges who had not previously registered for the course but wished to access JudgeGPT, which was not available to them. An additional 205 judges expressed interest in participating in the course to gain access to JudgeGPT. To manage this surge of interest—which could enhance our statistical power—while preserving the study's integrity, we decided against simply adding these new applicants to our control group, namely to our control Batch 2, as the newly interested judges had not been randomly selected. Therefore, we concluded that a second randomization was necessary to uphold the experiment's integrity and to increase the study's statistical power, allowing for the inclusion of 1,185 judges instead of the originally registered 980.

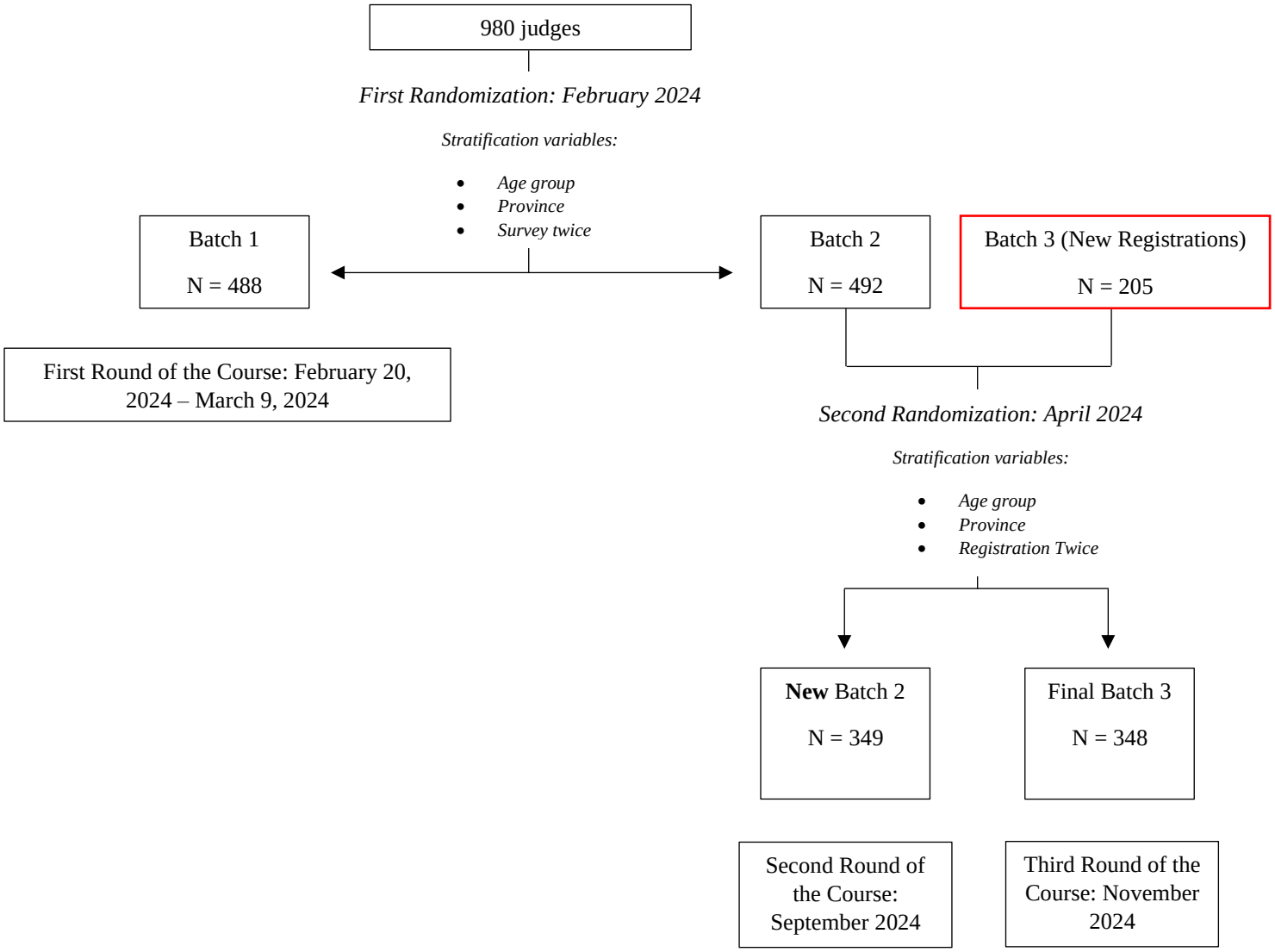
The second stage of randomization, therefore, took place on April 26, 2024. We first merged the original Batch 2 (492 judges) with Batch 3 of new registrations (n = 205).

These new 697 judges were divided into New Batch 2 (n =349), who will take the course in September 2024, and Final Batch 3 (n = 348), who will take the course in November 2024 and will serve as our control group.

Stratification in all instances was based on the province where the judge's court is located, the judge's age, and whether the judge participated in the survey more than once, which captured the judges' interest in the course. This approach ensured that we could detect the treatment effect among judges from all provinces and that they were comparable in age.

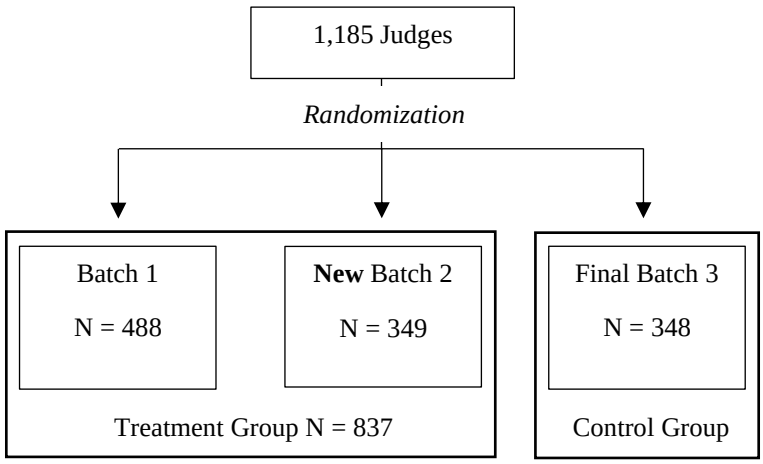
Therefore, to summarize the following figure below illustrates the old and new groups of judges:

Figure 1: Flow Chart of the Experimental Design




*Notes:* This figure shows the months over which randomization and training were conducted. The study originally randomized 980 consenting judges into treatment (488) and control (492) groups in February 2024. Due to high interest, 205 additional judges were recruited and randomized again in April 2024, resulting in the new Batch 2 (349) and Final Batch 3 (348) control groups. Stratified randomization is based on province, age, and survey response frequency.

Figure 2: Randomization into three Groups



## Stratification Variables are Listed Below

First Randomization

| Variable | Description |
|---|---|
| ***Timestamp*** | The last time the survey was conducted by a specific judge. |
| ***First Survey*** | The first time the survey was conducted by a specific judge. |
| ***Province*** | A categorical variable that is based on the administrative units of Pakistan, with 6 unique values:<br><br>Azad-Kashmir-Gilgit-Baltistan, Balochistan, Islamabad (federal territory), Khyber Pakhtunkhwa, Punjab, Sindh.<br><br>Azad-Kashmir and Gilgit-Baltistan are combined into one category due to the small sample for Gilgit-Baltistan ( + they share a common boundary). |
| ***Age group*** | A categorical variable that is based on the age of the judges, with 3 unique values: «<40», «40-49», «>=50». |
| ***Survey twice*** | This is a dummy variable that switches to 1 when Timestamp is not equal to First Survey. This variable represents the involvement of judges. |
| ***Block*** | We stratified the entire study population into subgroups with the same characteristics based on Province (6 unique values), Age group (3 unique values) and Survey twice (2 unique values). All judges are divided into 2*3*6 = 36 blocks. |

Second Randomization

| Variable | Description |
|---|---|
| ***Province*** | Stays the same. |
| ***Age group*** | A categorical variable that is based on the age of the judges, with 3 unique values: «<41», «41-47», «>=48». |
| ***Batch number*** | A categorical variable that is based on the date when the survey was completed, if the judge passed the survey after the first randomization, then he or she is determined in batch 3. Batch 2 is the result of the first stage of randomization. |
| ***Block*** | We stratified the entire study population into subgroups with the same characteristics based on Province (6 unique values), Age group (3 unique values) and Batch number (2 unique values). All judges are divided into 2*3*6 = 36 blocks. |

Table 1: Balance over Judge Characteristics

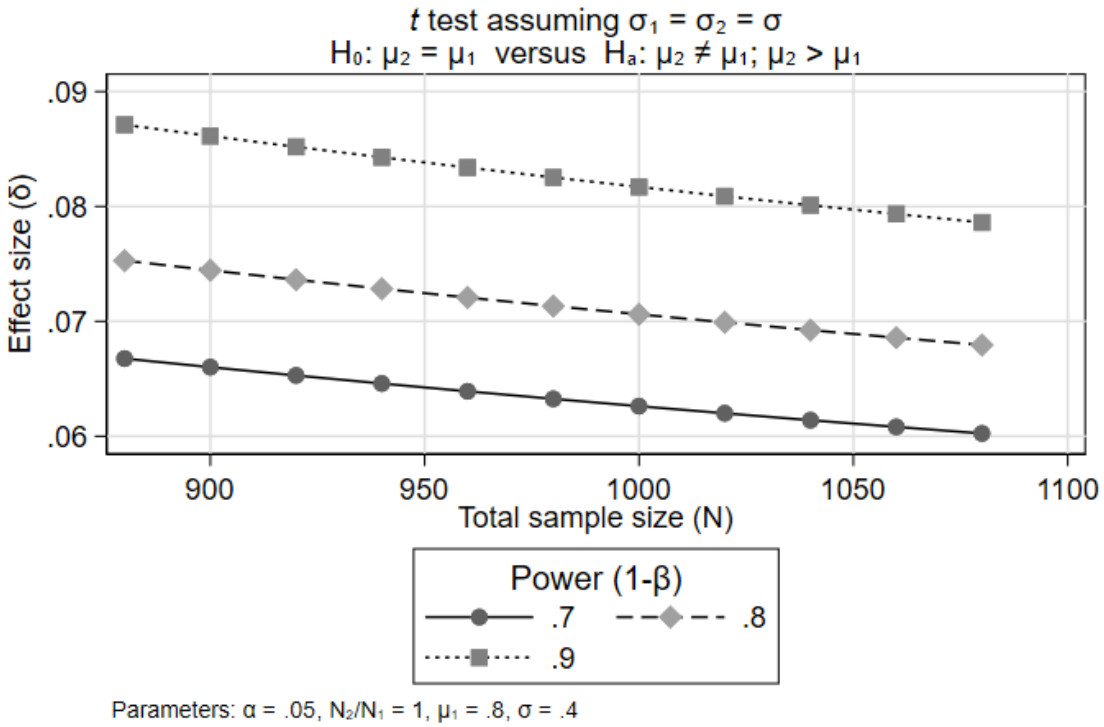
| | (1) | | (2) | | (3) | | (1)-(2) | | (1)-(3) | | (2)-(3) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Batch 1* | | *Final Batch 2* | | *Final Batch3* | | *Pairwise t-test* | | *Pairwise t-test* | | *Pairwise t-test* | |
| *Variable* | *N* | *Mean/SE* | *N* | *Mean/(SE)* | *N* | *Mean/(SE)* | *N* | *P-value* | *N* | *P-value* | *N* | *P-value* | *Total N* |
| *Age* | 488 | 42.531 (0.331) | 349 | 42.745 (0.403) | 348 | 42.828 (0.399) | 837 | 0.682 | 836 | 0.568 | 697 | 0.884 | 1,185 |
| *Gender* | 488 | 1.779 (0.019) | 349 | 1.762 (0.023) | 348 | 1.830 (0.020) | 837 | 0.579 | 836 | 0.062 | 697 | 0.025 | 1,185 |
| *Years of Experience* | 488 | 11.455 (0.321) | 349 | 11.892 (0.406) | 348 | 11.454 (0.401) | 837 | 0.399 | 836 | 0.999 | 697 | 0.443 | 1,185 |
| *AI Support* | 488 | 3.414 (0.038) | 349 | 3.438 (0.042) | 348 | 3.437 (0.045) | 837 | 0.665 | 836 | 0.697 | 697 | 0.979 | 1,185 |
| *Income* | 488 | 2.945 (0.030) | 349 | 2.968 (0.032) | 348 | 2.954 (0.038) | 837 | 0.584 | 836 | 0.845 | 697 | 0.769 | 1,185 |
| *Technology Experience* | 488 | 2.545 (0.031) | 349 | 2.593 (0.034) | 348 | 2.549 (0.037) | 837 | 0.300 | 836 | 0.938 | 697 | 0.379 | 1,185 |
| *Use of Online Legal Resources* | 488 | 2.973 (0.046) | 349 | 2.885 (0.056) | 348 | 2.879 (0.056) | 837 | 0.229 | 836 | 0.198 | 697 | 0.939 | 1,185 |
| *Number of Cases on the Desk* | 488 | 533.223 (24.446) | 349 | 600.570 (73.229) | 348 | 501.224 (28.864) | 837 | 0.383 | 836 | 0.398 | 697 | 0.207 | 1,185 |
| *Number of Decided Cases* | 488 | 115.971 (6.233) | 349 | 109.301 (6.224) | 348 | 112.017 (6.870) | 837 | 0.449 | 836 | 0.670 | 697 | 0.770 | 1,185 |
| *Number of Cases Concurrently Managed* | 488 | 158.768 (19.445) | 349 | 118.206 (12.821) | 348 | 144.299 (18.560) | 837 | 0.082 | 836 | 0.591 | 697 | 0.248 | 1,185 |
| *Hours spent on Legal Research and Writing Judgments* | 488 | 16.818 (1.144) | 349 | 17.656 (1.556) | 348 | 17.138 (0.685) | 837 | 0.664 | 836 | 0.810 | 697 | 0.761 | 1,185 |
| *Hours spent on Administrative Work* | 488 | 11.693 (0.496) | 349 | 13.149 (1.099) | 348 | 11.365 (0.624) | 837 | 0.227 | 836 | 0.681 | 697 | 0.159 | 1,185 |
| *Workload* | 488 | 6.475 (0.100) | 349 | 6.599 (0.109) | 348 | 6.279 (0.120) | 837 | 0.403 | 836 | 0.206 | 697 | 0.048 | 1,185 |
| *Work/Life Balance* | 488 | 5.818 (0.094) | 349 | 5.762 (0.113) | 348 | 5.784 (0.112) | 837 | 0.705 | 836 | 0.821 | 697 | 0.889 | 1,185 |
| *Confidence in Legal Research Abilities* | 488 | 6.721 (0.097) | 349 | 6.421 (0.112) | 348 | 6.624 (0.106) | 837 | 0.043 | 836 | 0.498 | 697 | 0.190 | 1,185 |
| *Confidence in Legal Writing Abilities* | 488 | 7.375 (0.084) | 349 | 7.103 (0.102) | 348 | 7.270 (0.096) | 837 | 0.040 | 836 | 0.411 | 697 | 0.234 | 1,185 |
| *Confidence in the Public Appearance* | 488 | 7.627 (0.092) | 349 | 7.450 (0.106) | 348 | 7.641 (0.107) | 837 | 0.207 | 836 | 0.923 | 697 | 0.205 | 1,185 |
| *Confidence in Administrative Work* | 488 | 7.516 (0.088) | 349 | 7.458 (0.094) | 348 | 7.652 (0.098) | 837 | 0.652 | 836 | 0.302 | 697 | 0.154 | 1,185 |
| *Expectations from AI for Judges* | 488 | 3.717 (0.028) | 349 | 3.731 (0.034) | 348 | 3.727 (0.030) | 837 | 0.757 | 836 | 0.811 | 697 | 0.936 | 1,185 |

*Notes:* 837 is the number of judges in Batch 1 and Final Batch 2, 836 is the number of judges in Batch 1 Final Batch 3, and 697 is the number of judges in Final Batch 2 and Final Batch 3. Judges were asked to rate their confidence in various aspects of their work on a scale of 1 to 10. Based on the answers, the variables «Confidence in legal research abilities», «Confidence in legal writing abilities», «Confidence in the public appearance» and «Confidence in administrative work» were formed. «Workload» and «Work/Life Balance» were also rated by the judges on a 10-point scale. «AI Support», «Expectations from AI for Judges» and «Use of Online Legal Resources» are assessed by judges on a 5-point scale. «Age», «Years of Experience», «Number of Cases on the Desk», «Number of Decided Cases» (value for last month), «Number of Cases concurrently managed», «Hours spent on Legal Research and Writing Judgments» (hours per week), «Hours spent on Administrative work» (hours per week) are quantitative variables. «Gender» is a categorical variable that is encoded and takes the value 1 if the judge is female, the value 2 if the judge is male otherwise 3. «Income» and «Technology Experience» are categorical variables that are encoded and take the value of 1 for Low, the value of 2 for Medium, the value of 3 for High otherwise 0.

**Power Analysis**

By analyzing preliminary data on course registrations of judges, we can determine the detectable effect size with our current sample of judges. For instance, with a power of 0.8, the minimum detectable effect (MDE) of our initial treatment using JudgeGPT is estimated to be below 0.10. Given our substantial sample of approximately 1000 judges, we have the capacity to identify effect sizes as small as 0.7 standard deviations. Figure 3 below illustrates how our Minimum Detectable Effect (MDE) varies with sample size at different power levels.

Figure 3: Power Analysis for Course Participation




*Notes:* This graph represents effect size for a two-sample means test. On the X-axis, the sample size, the test is performed taking into account the difference in the size of the control group (Batch 2) and the treatment group (Batch 1) in this case, the ratio N2/N1 = 492/488. On the Y-axis, the estimated effect size for participation, participation is dummy variable that switches to 1 if the judge participated in at least one lecture and 0 otherwise.

## Appendix Instructions Detail from Raw Registration File to Final Sample

1) There were 2,063 responses in the raw data for registration file with many repeat responses.

2) Originally registered Judges in February 2024 had 1,798 registrations by judges (including duplicate registrations).

3) We checked by email, name and birth date that the judges are not in batch 2 and batch 1 to find the true unique new registrations. This gave us 205 new judges who registered for the course.

4) 980 were the group of judges that originally registered in February 2024, without new registrations of 205.

5) Total sample of judges randomized into three batches are 1185 that includes new registrations.