30th October 2024

Courts of Tomorrow

Experimental Design

This analysis plan outlines the methodology of our randomization scheme for judges in Pakistan. The random assignment of judges was conducted in two distinct waves for registration into JudgeGPT subscriptions.

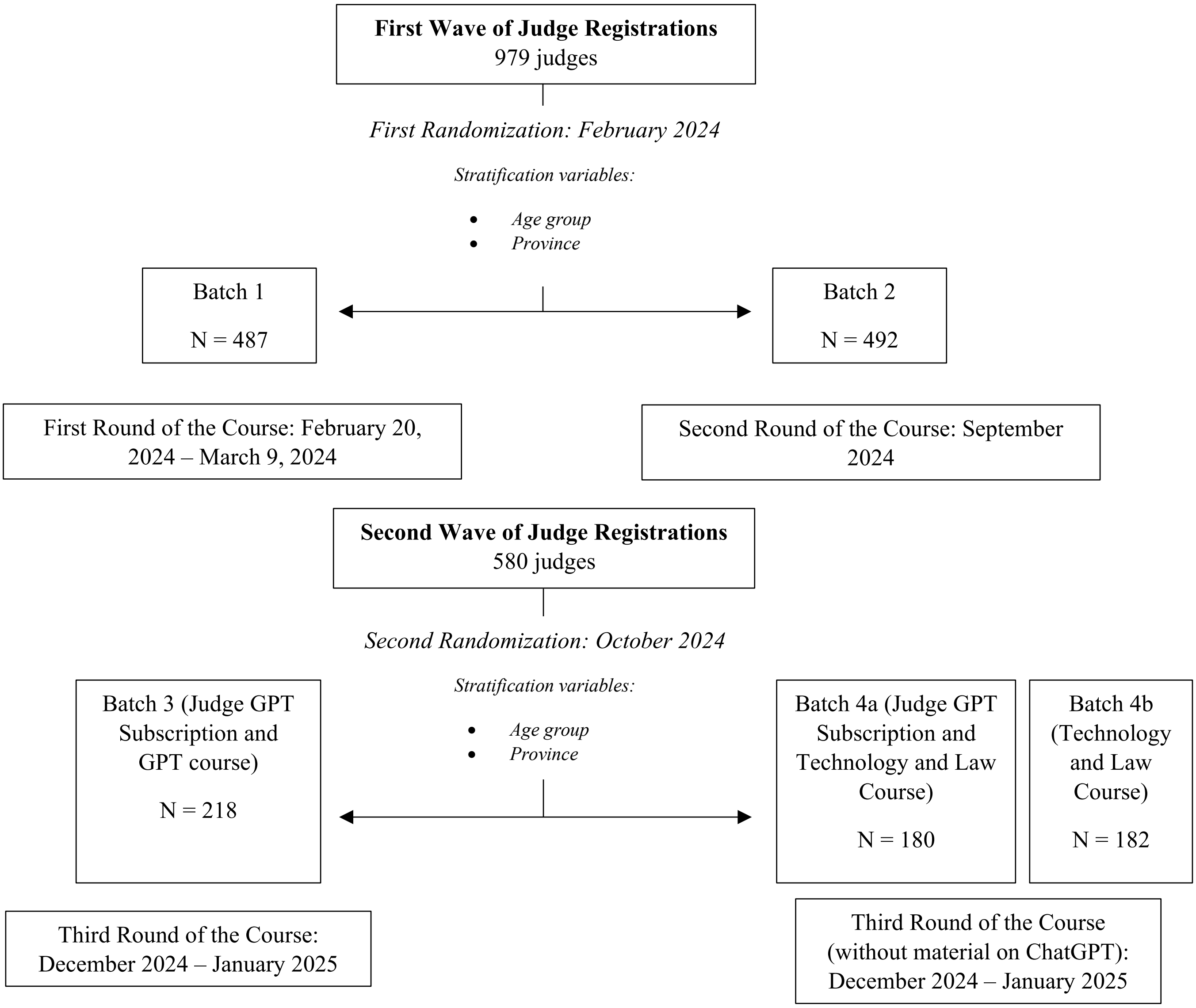
In February 2024, the first wave of registration saw 979 judges from Pakistan's lower courts sign up to participate in our experiment. We randomly assigned these 979 judges into two groups: 487 judges were allocated to the treatment group (Batch 1) and provided with access to the JudgeGPT subscription and GPT instruction course, while the remaining 492 judges were designated as the control group (Batch 2), scheduled to receive the same access in September 2024. This setup allows for a randomized control trial comparing the outcomes of Batch 1 and Batch 2. Following the initial random assignment, the introduction of the password-protected JudgeGPT, designed specifically to prevent spillovers, sparked considerable interest among judges who had not initially registered for the course but were nonetheless eager to participate but could not access GPT or access the course. An additional 580 judges expressed interest in the course and the JudgeGPT tool. To preserve the study's integrity, we decided against simply adding these new applicants to our control group (Batch 2), as they were not randomly assigned. Therefore, a second randomization was conducted to maintain the integrity of the study and increase its statistical power, accommodating a total of 1559 judges instead of the initially registered 979. This means more than 50% of the trial court judges (court of first instance) in Pakistan registered to participate in our experiment.

In October, the second wave of randomization, therefore, took place on October 23, 2024 for 580 judges. The 580 judges were randomly assigned into Batch 3 (n = 218), who will take the course in December 2024 and January 2025, and Batch 4 (n = 362). Batch 3 judges would get the same treatment as Batch 1 and 2: JudgeGPT course and JudgeGPT subscription.

Batch 4, however, is further randomized into two subgroups: Batch 4a and Batch 4b. Batch 4a is randomly assigned to receive JudgeGPT training and a placebo course on Technology and Law in December 2024 and January 2025, along with a GPT subscription (and an anti-hallucination warning in GPT). Batch 4b will also take the generic Technology and Law course during the same period but will not receive a GPT subscription. The key difference is that Batch 4a will have access to the GPT subscription with a hallucination warning, while Batch 4b will not. Both groups, however, will attend the Generic Law and Technology classes at the same time that Batch 3 is receiving the JudgeGPT course. This will allow us to assess the impact of access to GPT tools on judges' learning and decision-making. For a summary of this design, please see Figure 1 on the next page.

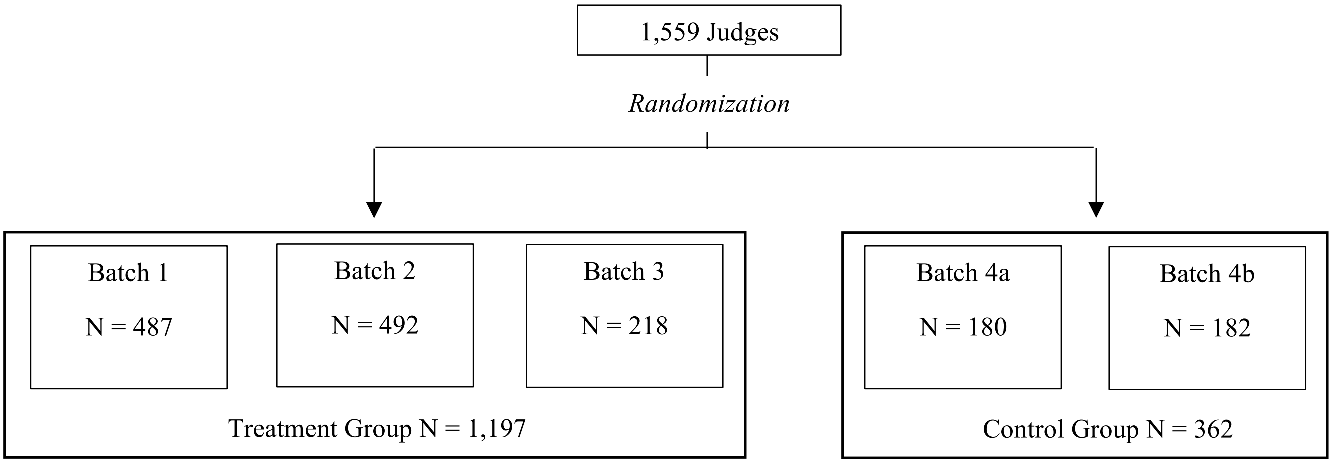
Stratification in all instances were based on the province in which the judge's court is located, the age of the judge and whether the judge participated in the survey more than once that captured interest of the course by the judges. This was done so we are able to detect treatment effect by judges in all provinces and they are similar in age. The following flow chart summarizes the main design of the experiment with 2 waves of judges:

Figure 1: Flow Chart of the Experimental Design


First Wave of Judge Registrations
979 judges
First Randomization: February 2024
Stratification variables:
Age group
Province
Batch 1
N = 487
Batch 2
N = 492
First Round of the Course: February 20, 2024 – March 9, 2024
Second Round of the Course: September 2024
Second Wave of Judge Registrations
580 judges
Second Randomization: October 2024
Stratification variables:
Age group
Province
Batch 3 (Judge GPT Subscription and GPT course)
N = 218
Batch 4a (Judge GPT Subscription and Technology and Law Course)
N = 180
Batch 4b (Technology and Law Course)
N = 182
Third Round of the Course: December 2024 – January 2025
Third Round of the Course (without material on ChatGPT): December 2024 – January 2025

Notes: This figure shows the months over which randomization and training were conducted. The study randomized 979 consenting judges into treatment (487) and control (492) groups in February 2024. Due to high interest, 580 additional judges were recruited and randomized again in October 2024, resulting in Batch 3 (218), Batch 4a (180) and Batch 4b (182). Stratified randomization is based on province, age, and survey response frequency to ensure similar treatment and control groups.

Figure 2: Randomization into four 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: Azad-Kashmir-Gilgit-Baltistan, Balochistan, Islamabad (federal territory), Khyber Pakhtunkhwa, Punjab, Sindh. 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: «<40», «40-46», «>=47».
Group number	The variable takes on two values, based on the date the form was completed. If a judge completed the survey before September, it takes the value 1, and 2 otherwise.
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 Table

	(1)	(2)	(3)	(4)	(5)	(1)-(2)	(1)-(3)	(1)-(4)	(1)-(5)	(2)-(3)	(2)-(4)	(2)-(5)	(3)-(4)	(3)-(5)	(4)-(5)
						Pairwise	Pairwise	Pairwise	Pairwise	Pairwise	Pairwise	Pairwise	Pairwise	Pairwise	Pairwise
	Batch 1	Batch 2	Batch 3	Batch 4a	Batch 4b	t-test	t-test	t-test	t-test	t-test	t-test	t-test	t-test	t-test	t-test
Variable	Mean/(SE)	Mean/(SE)	Mean/(SE)	Mean/(SE)	Mean/(SE)	P-value	P-value	P-value	P-value	P-value	P-value	P-value	P-value	P-value	P-value
Age	42.520	42.963	42.417	42.800	42.451	0.337	0.855	0.645	0.909	0.324	0.786	0.393	0.574	0.961	0.627
	(0.332)	(0.321)	(0.450)	(0.510)	(0.507)										
Gender	1.778	1.801	1.775	1.822	1.824	0.390	0.930	0.200	0.178	0.446	0.526	0.486	0.244	0.222	0.961
	(0.019)	(0.018)	(0.028)	(0.029)	(0.028)										
Years of Experience	11.446	11.614	11.379	11.583	11.774	0.716	0.903	0.825	0.623	0.672	0.960	0.812	0.769	0.591	0.809
	(0.322)	(0.331)	(0.447)	(0.531)	(0.584)										
AI Support	3.413	3.453	3.413	3.406	3.440	0.441	0.999	0.922	0.708	0.543	0.512	0.848	0.931	0.746	0.698
	(0.038)	(0.037)	(0.055)	(0.063)	(0.061)										
Income	2.947	2.915	3.069	2.950	3.027	0.447	0.015**	0.954	0.177	0.002***	0.543	0.060*	0.066*	0.531	0.283
	(0.030)	(0.030)	(0.041)	(0.050)	(0.052)										
Technology Experience	2.544	2.569	2.596	2.600	2.555	0.568	0.302	0.308	0.848	0.587	0.570	0.801	0.951	0.502	0.489
	(0.031)	(0.031)	(0.040)	(0.045)	(0.047)										
Use of Online Legal Resources	2.971	2.923	2.748	2.678	2.736	0.464	0.014**	0.001***	0.012**	0.055*	0.007***	0.046**	0.526	0.919	0.602
	(0.046)	(0.047)	(0.078)	(0.078)	(0.080)										
Number of Cases on the Desk	532.456	515.085	524.936	654.539	474.165	0.615	0.861	0.384	0.193	0.818	0.320	0.359	0.364	0.325	0.208
	(24.484)	(24.322)	(35.337)	(138.142)	(37.452)										
Number of Decided Cases	115.624	116.222	126.587	106.933	103.874	0.944	0.489	0.368	0.258	0.508	0.321	0.223	0.229	0.176	0.783
	(6.236)	(5.782)	(14.550)	(7.367)	(8.319)										
Number of Cases Concurrently	158.992	130.701	126.404	174.989	118.077	0.209	0.191	0.712	0.117	0.823	0.272	0.541	0.245	0.721	0.181
Managed	(19.484)	(11.228)	(15.539)	(38.760)	(17.334)										
Hours spent on Legal Research and	16.809	17.766	16.179	15.711	17.286	0.554	0.650	0.444	0.757	0.253	0.152	0.755	0.689	0.393	0.242
Writing Judgments	(1.146)	(1.142)	(0.786)	(0.864)	(1.030)										
Hours spent on Administrative Work	11.696	12.055	11.202	12.883	11.236	0.704	0.589	0.319	0.625	0.444	0.539	0.470	0.206	0.975	0.222
	(0.497)	(0.804)	(0.769)	(1.082)	(0.798)										
Workload	6.472	6.472	5.940	6.078	6.198	0.996	0.002***	0.037**	0.140	0.002***	0.037**	0.141	0.522	0.225	0.593
	(0.100)	(0.100)	(0.142)	(0.161)	(0.157)										
Work/Life Balance	5.815	5.799	5.550	5.683	5.714	0.903	0.131	0.470	0.570	0.160	0.531	0.637	0.538	0.438	0.887
	(0.094)	(0.097)	(0.148)	(0.157)	(0.151)										
Confidence in Legal Research Abilities	6.723	6.650	6.220	6.150	6.231	0.593	0.003***	0.001***	0.009***	0.011**	0.005***	0.024**	0.733	0.960	0.714
	(0.098)	(0.094)	(0.140)	(0.150)	(0.161)										
Confidence in Legal Writing Abilities	7.382	7.321	6.927	6.833	7.099	0.606	0.002***	0.002***	0.076*	0.007***	0.005***	0.161	0.629	0.342	0.190
	(0.084)	(0.083)	(0.121)	(0.151)	(0.135)										
Confidence in the Public Appearance	7.634	7.616	7.289	7.311	7.505	0.884	0.039**	0.084*	0.474	0.049**	0.101	0.537	0.918	0.300	0.387
	(0.092)	(0.089)	(0.140)	(0.163)	(0.155)										
Confidence in Administrative Work	7.520	7.596	7.344	7.400	7.495	0.528	0.256	0.475	0.881	0.097*	0.235	0.539	0.769	0.431	0.639
	(0.088)	(0.082)	(0.127)	(0.142)	(0.143)										
Expectations from AI for Judges	3.717	3.738	3.697	3.739	3.720	0.580	0.718	0.672	0.953	0.445	0.983	0.733	0.517	0.728	0.765
	(0.028)	(0.026)	(0.046)	(0.045)	(0.046)										
Observations	487	492	218	180	182	979	705	667	669	710	672	674	398	400	362

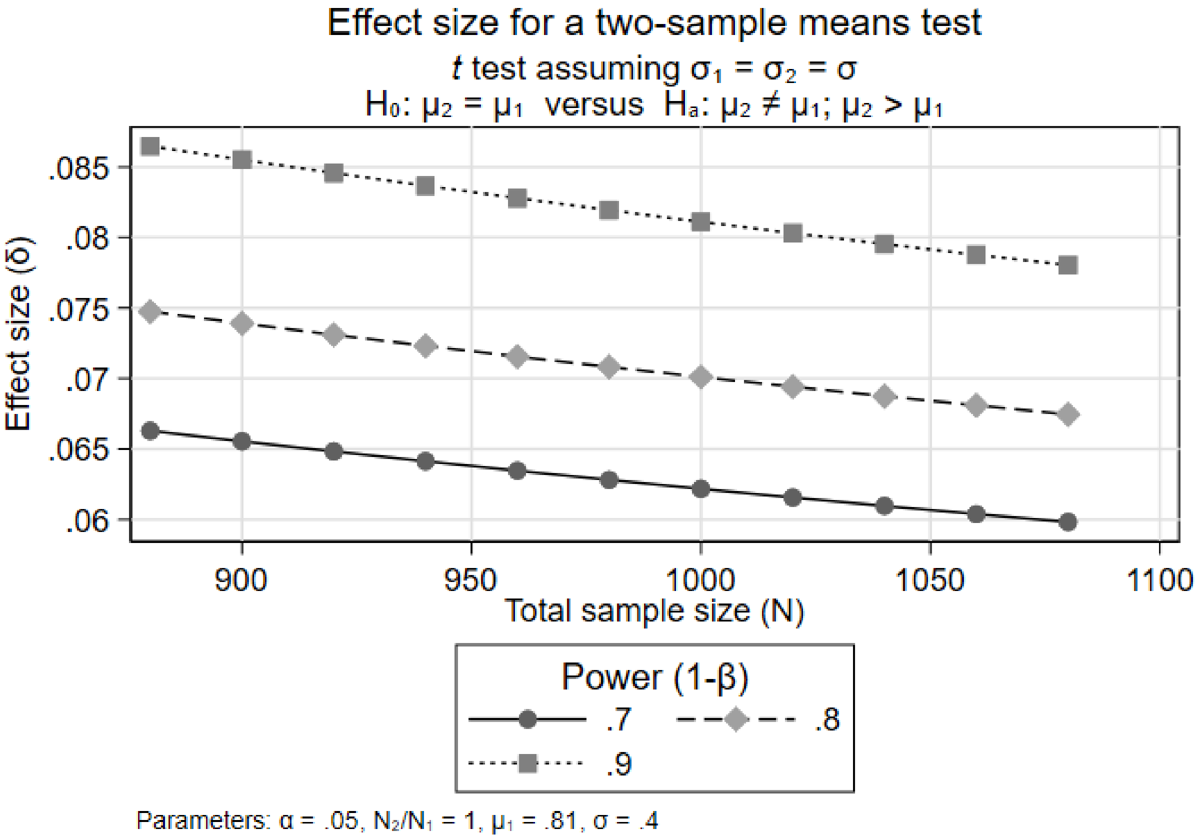
Notes: Judges were asked to rate their confidence in various aspects of their work on a scale of 1 to 10. Based on the answers, the variables «Confidence in legal research abilities», «Confidence in legal writing abilities», «Confidence in the public appearance» and «Confidence in administrative work» 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 4-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.

Table 2: Estimated Effect Size and Group Mean

alpha	power	N	N1	N2	delta	m1	m2	sd
.05	.7	979	487	492	0.063	0.807	0.870	0.395
.05	.8	979	487	492	0.071	0.807	0.878	0.395
.05	.9	979	487	492	0.082	0.807	0.889	0.395

Notes: this table represents effect size for a two-sample means test. Alpha is significance level. N is the total sample size, N1 is the size of the treatment group (Batch 1), and N2 is the size of the control group (Batch 2). Delta is the estimated effect size. m1 is the mean of participation in treatment group, where participation is dummy variable that switches to 1 if the judge participated in at least one lecture and 0 otherwise. m2 is the estimated mean of participation in control group.

Figure 3: Effect size for a two-sample means test for Batch 1 and Batch 2



Notes: this graph represent 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. 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 2,962 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 580 new judges who registered for the course.

4) 979 were the group of judges that originally registered in February 2024, without new registrations of 580.

5) Total sample of judges randomized into four batches are 1,559 that includes new registrations.