# **Top Candidate Notifications Experiment Report**

Akram Assaf, Dan Begley-Groth, Moonsoo Kim

W241

Spring 2025

## **Introduction and Background**

Job seekers often struggle to gauge their competitiveness for positions, leading them to apply for jobs they are unlikely to land. On Bayt.com (a leading Middle Eastern job site), user surveys revealed frustration at the lack of feedback after applying and a desire for clarity on their chances​. Prior research suggests that providing feedback on application ranking could encourage more targeted applications and higher engagement​. In this context, we ask: **What is the causal impact on job application behavior of notifying candidates that they are a “top candidate” for a job?** This experiment tests the intervention, which informs users of their relative rank (e.g., “You are ranked X out of Y applicants”) for specific job postings. By framing the problem in terms of **potential outcomes** under treatment vs. control, we use a randomized controlled trial (RCT) to identify the causal effect of these notifications on user behavior.

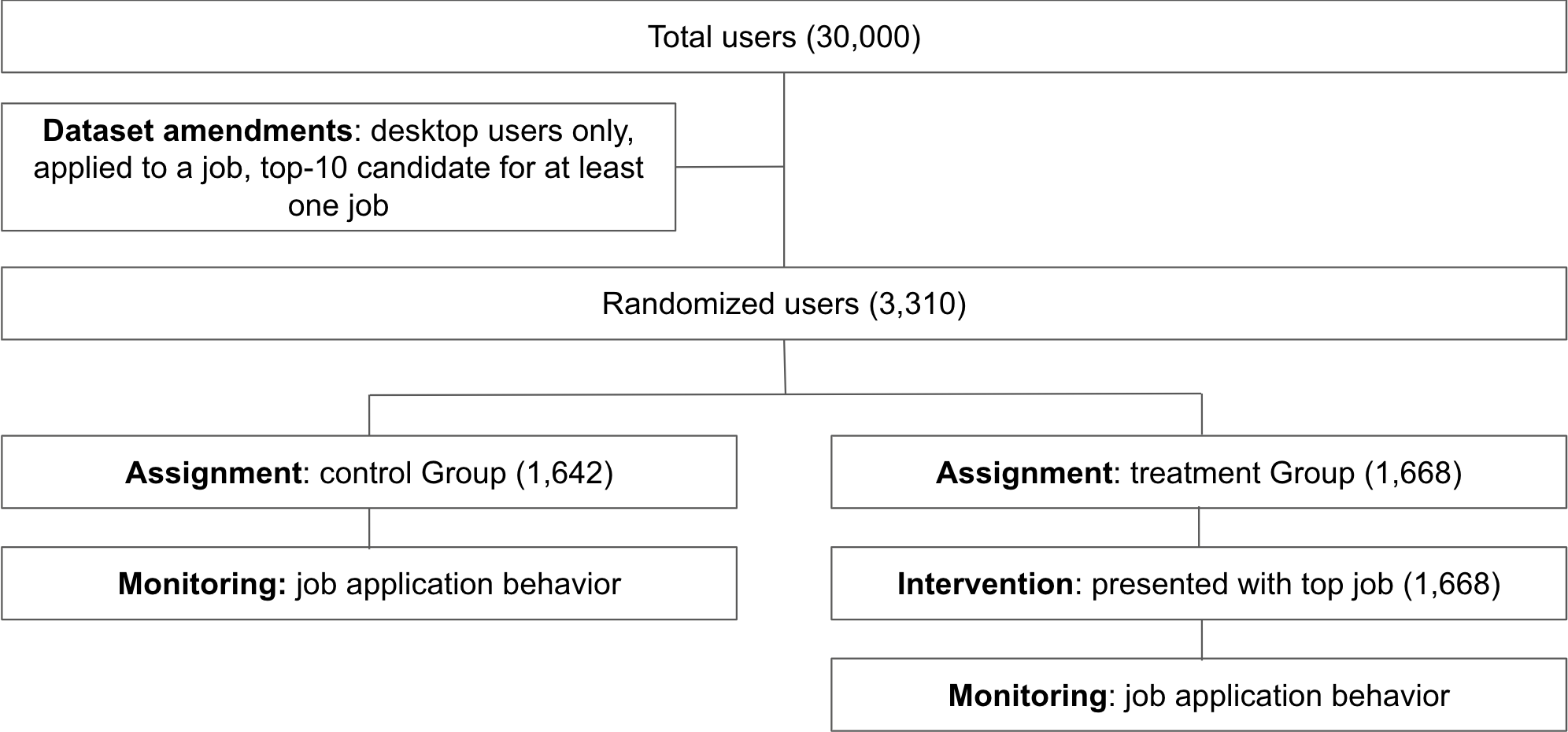
## **Hypothesis**

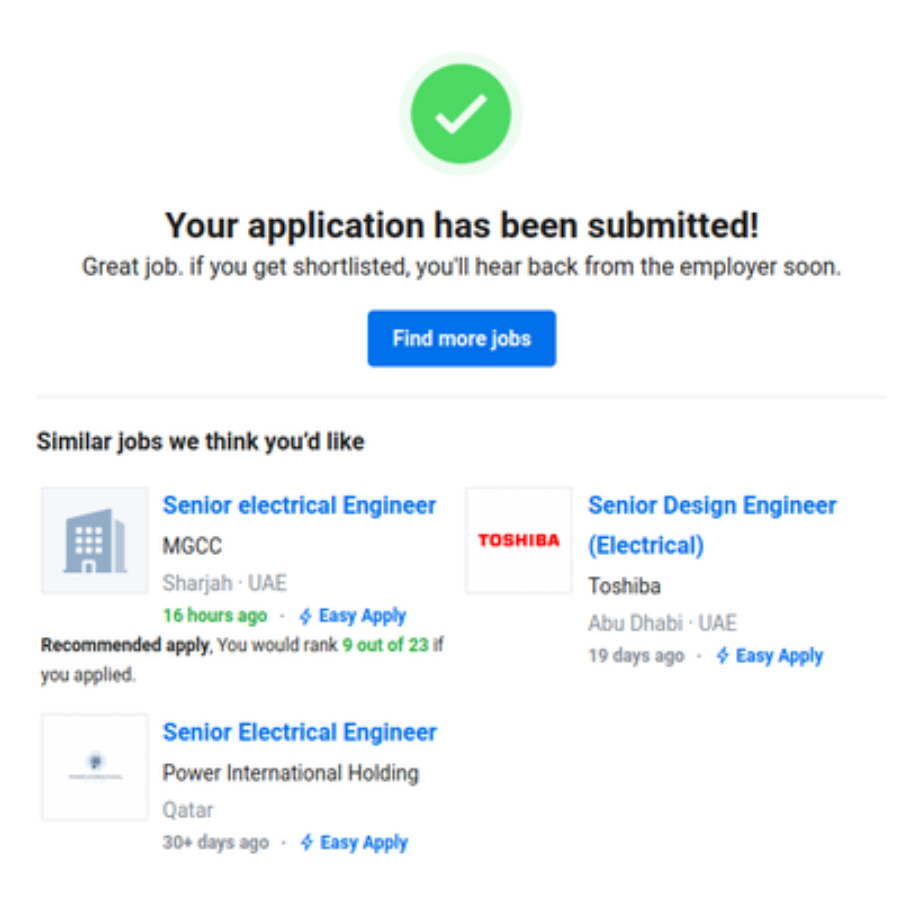
We hypothesized that providing job seekers with feedback about being a top candidate would significantly change their behavior. **H1:** Users shown jobs for which they are a top candidate will apply to those jobs at a much higher rate than users not shown such information​. In other words, the intervention was intended to increase applications to “top candidate” jobs *without* reducing applications to different jobs, thereby raising the overall application count. The null hypothesis, in contrast, is that the application behavior would not differ between the treatment and control groups, after accounting for any general trends. We also anticipated possible **heterogeneous treatment effects (HTEs)** by the strength of a user’s ranking: for example, users barely in the top 10 might respond differently than users ranked at the very top, and both of these cases could vary further by the total applicants in the job inbox.

## **Experiment Design**

**Design framework:** We implemented a between-subjects randomized experiment following an R→O→X→O design (ROXO)​. First, eligible users were **Randomized (R)** into either treatment or control. We observed a **pre-treatment baseline (O)** for each user’s job search activity (using historical logs of job views and applications)​. The treatment group then received the **Intervention (X)**: a notification interface highlighting jobs where the user’s applicant rank was in the top 10. After exposure, we **Observed (O)** each user’s post-intervention behavior in terms of job views and applications​. This design, by comparing outcomes before and after and between groups, aligns with the **potential outcomes framework**, in which each user has two potential post-intervention outcomes (with or without the notification). Random assignment allows us to estimate the average treatment effect by comparing the groups.

**Eligibility and participants:** To focus the test, we restricted it to desktop web users of Bayt.com who had applied to at least one job and who were a top-10 ranked candidate for at least one open job during the experiment period​. This ensured that every participant *could* be shown a “top candidate” job; otherwise, the treatment would be irrelevant or empty. Mobile users were excluded due to UI limitations that made it challenging to display notifications on small screens. In total, **3,312** users met these criteria. Of these, **3,310** were successfully randomized. Two records were dropped due to data issues because an internal testing user had multiple profiles. The randomized sample was split into **Treatment (1,668 users)** and **Control (1,642 users)**. Random assignment was implemented by distributing users based on their last digit (odd or even). Whenever a tagged user applied for a job on the site, the front-end would display a set of similar jobs and trigger a notification for one or more of these jobs if the user was in the treatment group. Conversely, no special notification was presented for control users. Users could not opt in or out of the notification when presented with the similar jobs modal, so **compliance was perfect**. All treated users received the prompt when appropriate, and no control users saw it. This means the **intent-to-treat equals the treatment effect on the treated** in our study, simplifying causal interpretation.

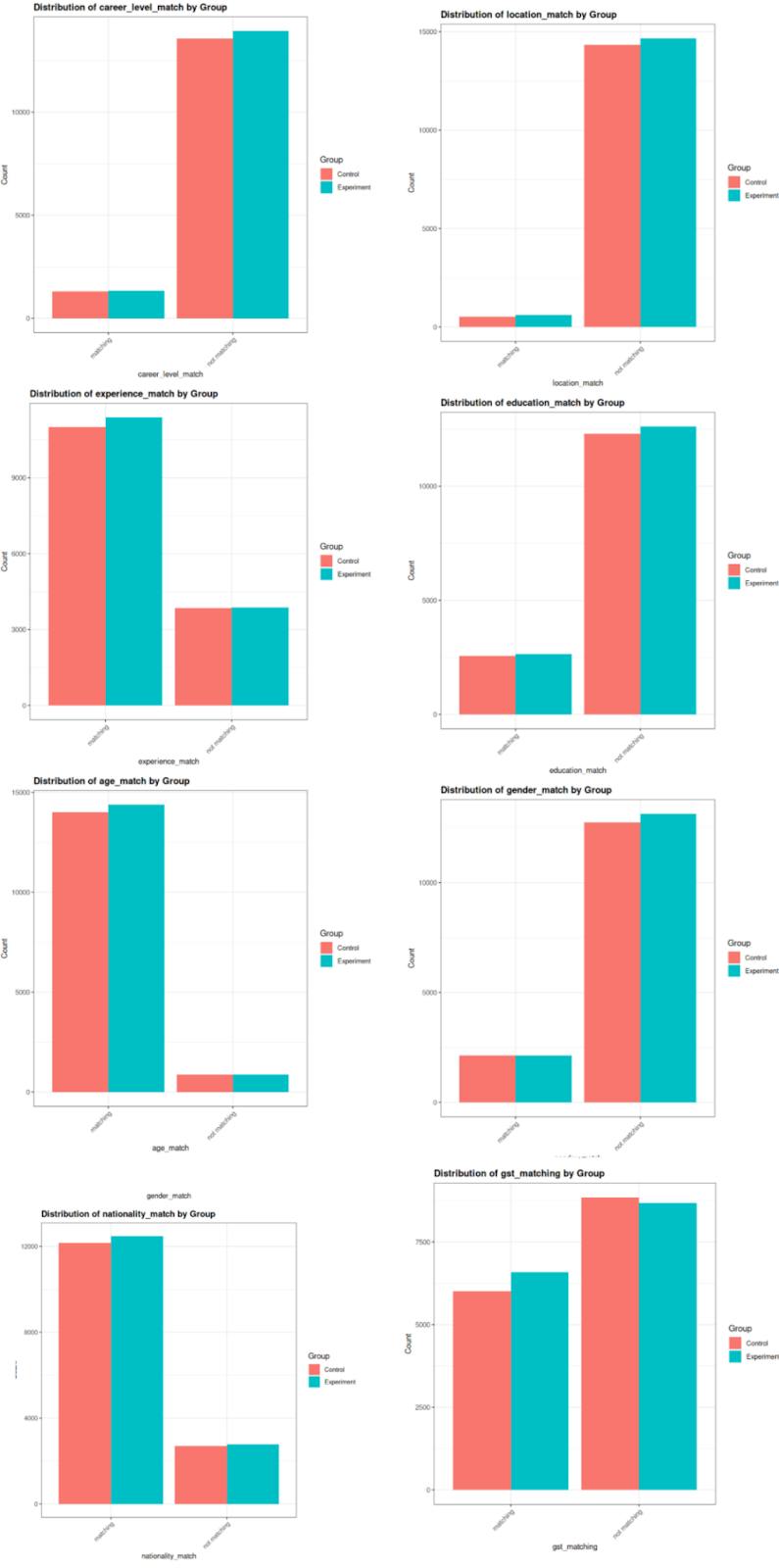
****

**Figure 1: Example of similar jobs modal with a treatment message**

**Treatment details:** For users in the treatment group, when applying for any job, a pop-up modal appears, informing them that they are a top candidate for other jobs. As shown in **Figure 1**, the modal lists these jobs, including their titles, links, and application ranks. For example, a user might see: *“Recommended apply. You would rank 3 of 23 if you applied”* The user could then click on these suggested jobs, dismiss the notification, or even view and possibly apply to a non-marked job. As they continued job searching in that session, the site would continue to show the list of top-candidate jobs for further browsing​. Users in the control group experienced the site normally with **no special notifications** for any of the jobs. Notably, they could still discover and apply to any jobs on their own, but they were not explicitly told about their “top candidate” status. The notification only considered jobs where the user’s rank was 10 or less.

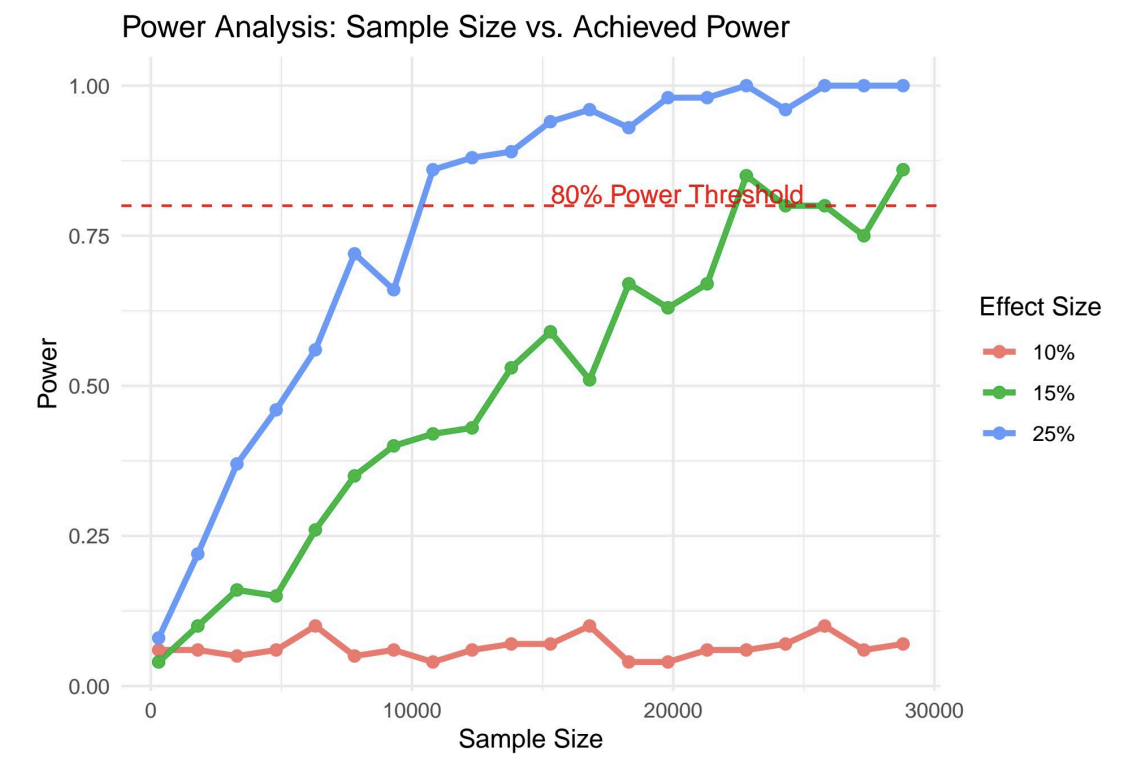
**Outcomes and measures:** The primary outcome of interest was user application behavior, measured in two main ways: (1) **Top-candidate application rate** – the proportion of users who applied to at least one “top candidate” job (as defined above) during the post-intervention period; and (2) **Top-candidate applications per user** – the average number of job applications to top-candidate jobs per user. These directly capture engagement with the recommended high-fit jobs. We also tracked **total job applications per user** and **job views per user** as secondary metrics, since an increase in targeted applications might translate to more overall applications and would likely be preceded by more job listing views (users clicking the suggestions). We designated job views as a *leading indicator* and applications as the *lagging metric* to validate that we have no unintended attrition due to technical issues. All metrics were recorded throughout the experiment window, which covered the period after users received the intervention, along with a parallel window for the control.

**Figure 2: User-Job fit covariates**



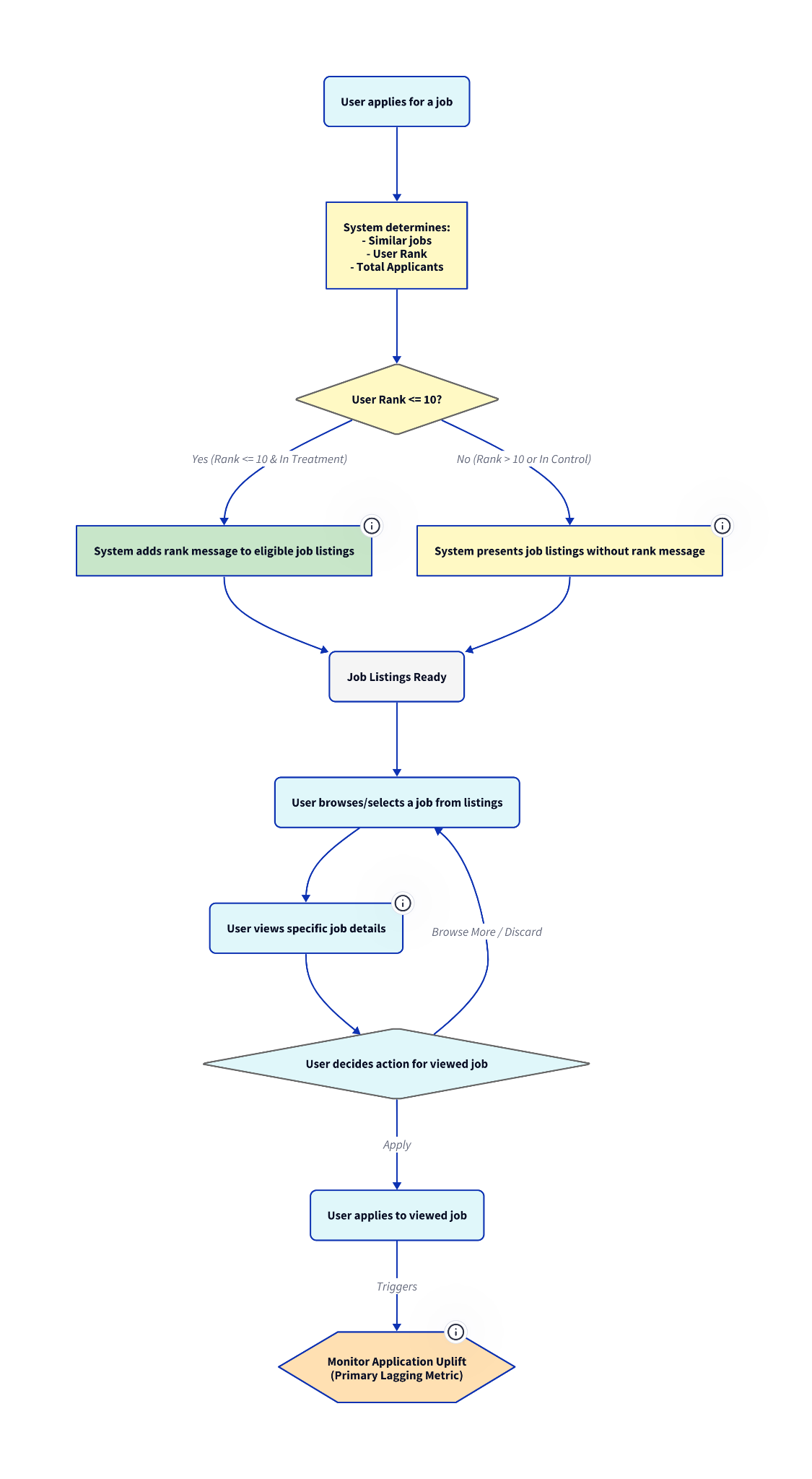
We note that the distribution of applications per user was highly skewed (many users made zero or few applications in or before the observation window, while a few applied to many jobs), which was considered in our analysis (using comparisons of means and regression models with robust standard errors). Additionally, several user- and job-specific **covariates** were collected to explore heterogeneity and adjust for any residual imbalances. These include user demographics and experience, as well as job “fit” indicators such as career level match, years of experience, location, education, etc. While randomization should ensure covariate balance in expectation, we collected these data to check balance and possibly increase the precision of estimates (Figure 2).

**Sample size and power:** The experiment’s sample of ~3.3k users was constrained by the number of active job seekers meeting the eligibility criteria in the timeframe. A power analysis indicated that this sample would likely be underpowered to detect small effects. Specifically, we estimated that **≈10,000 users** would be needed for 80% power to detect a +25% relative increase in the primary metric, and **≈25,000 users** would be needed to detect a +15% increase. Effects on the order of 10% or smaller would almost certainly not reach statistical significance with only a few thousand users. Thus, our experiment might only detect very large effects, and failure to reject the null could be due to insufficient power rather than the absence of any effect. This should be kept in mind when interpreting results.



**Figure 3: Power Analysis**

**Flow of participants:** In CONSORT-style terms, we can summarize the flow as follows. A total of 3,312 users were assessed and found eligible (desktop users with ≥1 top-candidate job). Two users were excluded due to technical issues before the assignment. **3,310 users** were randomized: 1,668 were allocated to **the Treatment** Group and 1,642 to the **Control Group**. All allocated participants were tracked in the outcome data. There was no attrition, as outcomes were collected via server database logs for all users, regardless of subsequent activity. Users who did not return to the site simply had zero post-intervention interactions but were still included in the analysis. We thus analyzed all 3,310 users in the groups to which they were assigned (an intention-to-treat analysis). The randomization produced treatment and control groups of approximately equal size and composition, as expected. We verified that there were no significant differences in pre-experiment application activity or key covariates between the groups. For example, treatment vs. control had virtually identical distributions of “top candidate” job counts (no significant difference, *p* = 0.18) and a negligible difference in sample size (*p* = 0.64)​. This confirms the random assignment was successful and any outcome differences can be causally attributed to the intervention (given perfect compliance and no contamination).

****

**Figure 4: Experiment Flow**

## 

## **Analysis Plan**

All user interactions (views and applications) were logged with timestamps, resulting in a rich dataset of 939,806 event records for this experiment. We aggregated this event data by user–job pair to calculate the total views and applications per job per user during and before the experiment. We then compared the experimental and control groups on these outcomes. Our analysis proceeded in several steps. First, we conducted **descriptive and preliminary analyses,** including checks of baseline activity (pre-experiment views and applications) and overall engagement changes before and after the intervention. This confirmed that both groups exhibited an increase in job search activity during the experiment period, likely due to external factors (e.g., active users usually apply to multiple jobs during a specific active job-hunting period), underscoring the importance of using the control group as a baseline​. Next, our primary analysis focused on **between-group differences during the experiment**, using two-sample *t*-tests and regression models to assess the treatment effect on views and applications. Because the distribution of views and applications per user–job was highly skewed (a few users clicked or applied to many jobs, far more than most), we applied an **outlier removal** procedure to avoid distortion. Specifically, within each subgroup analysis, we removed extremely high values using an interquartile range (IQR) filter, excluding user–job pairs with views or applications more than 1.5 times the IQR above the 75th percentile. This eliminated a small number of “power users” or potential bots with abnormally high activity counts, resulting in a more stable comparison. Finally, we employed **multiple regression** to adjust for the covariates described above. This allowed us to estimate the treatment effect on engagement while controlling for the quality of the match between user and job, as well as to examine which covariates independently predict engagement. Separate regression models were run for the outcomes of job views and job applications. We also stratified some analyses by job rank and competition level (e.g., examining only top-4 or top-10 ranked positions, and jobs with larger applicant pools) to see if the treatment impact varied in those contexts. All statistical tests used a significance level of α = .05, with key results reported with p-values, standard errors, and confidence intervals.

## **Results**

### **Overall Treatment Effect on Applications**

Contrary to our primary hypothesis, the experiment **did not detect a significant increase in job applications as a result of the top-candidate notifications**. Both the treatment and control groups showed a substantial rise in application activity in the post-intervention period. Still, **the increase was virtually the same in both groups**, indicating that the treatment had no additional effect.

**Table 1** illustrates this with the control group, where users increased from an average of about **0.83** applications per day pre-intervention to approximately **2.12** applications per day post-intervention (an increase of **~1.29**). Treated users experienced a nearly identical jump, from **~0.96** to **~2.07** per day (an increase of **~1.11**). In “Table 2”, the DiD regression’s interaction term was small and statistically insignificant (point estimate **~–0.17** applications/day, *p* ≈ **0.48**). In other words, **the treatment group did not show any improvement beyond that of the control group** during the experiment. We therefore cannot reject the null hypothesis – providing “You are a top candidate” feedback **did not materially change the overall number of job applications submitted**.

Several elements indicate that this null result is reliable. Initially, the absence of an effect was evident even in the basic post-period comparison, and this lack remained after accounting for users’ baseline activity. A strong overall time effect was observed: on average, all users became more active in job applications throughout the experimental period (likely due to the similar-jobs modal window after an application, clustering control and treatment records, with job application recency acting as a confounding factor, indicating a surge in job applications during specific period). This increase overshadowed any minor treatment effect. The DiD analysis in **Table 2** officially confirmed that the treatment’s additional impact was statistically indistinguishable from zero. It is also noteworthy that the point estimate for the treatment effect was slightly negative, suggesting that treated users applied slightly less than controls, although the difference was not significant. We do not view this as a legitimate negative effect but rather as evidence that there was certainly no substantial positive effect. Based on the confidence interval, we can confidently assert that the true average effect on applications was very limited (likely within ±10% of the baseline), which is below our detection threshold. Consequently, this outcome indicates that the Hypothesis was not supported: being notified of top-candidate status did not lead to a significantly higher application rate compared to users who were not notified.

**Table 1: Within‑group Pre → Post changes**

**(two regressions per group: one on views, one on applications)**

| **Metric** | **Group** | **Pre‑period mean (β₀)** | **Post shift Δ (β₁)** | **SE (Δ)** | **95 % CI** | **Sig** |
| --- | --- | --- | --- | --- | --- | --- |
| **Views / day** | **Control** | 2.43 | + 2.94 | 0.37 | 2.21 – 3.67 | \*\*\* |
| **Views / day** | **Experiment** | 2.45 | + 2.78 | 0.44 | 1.91 – 3.64 | \*\*\* |
| **Apps / day** | **Control** | 0.83 | + 1.29 | 0.16 | 0.98 – 1.59 | \*\*\* |
| **Apps / day** | **Experiment** | 0.96 | + 1.11 | 0.19 | 0.73 – 1.49 | \*\*\* |

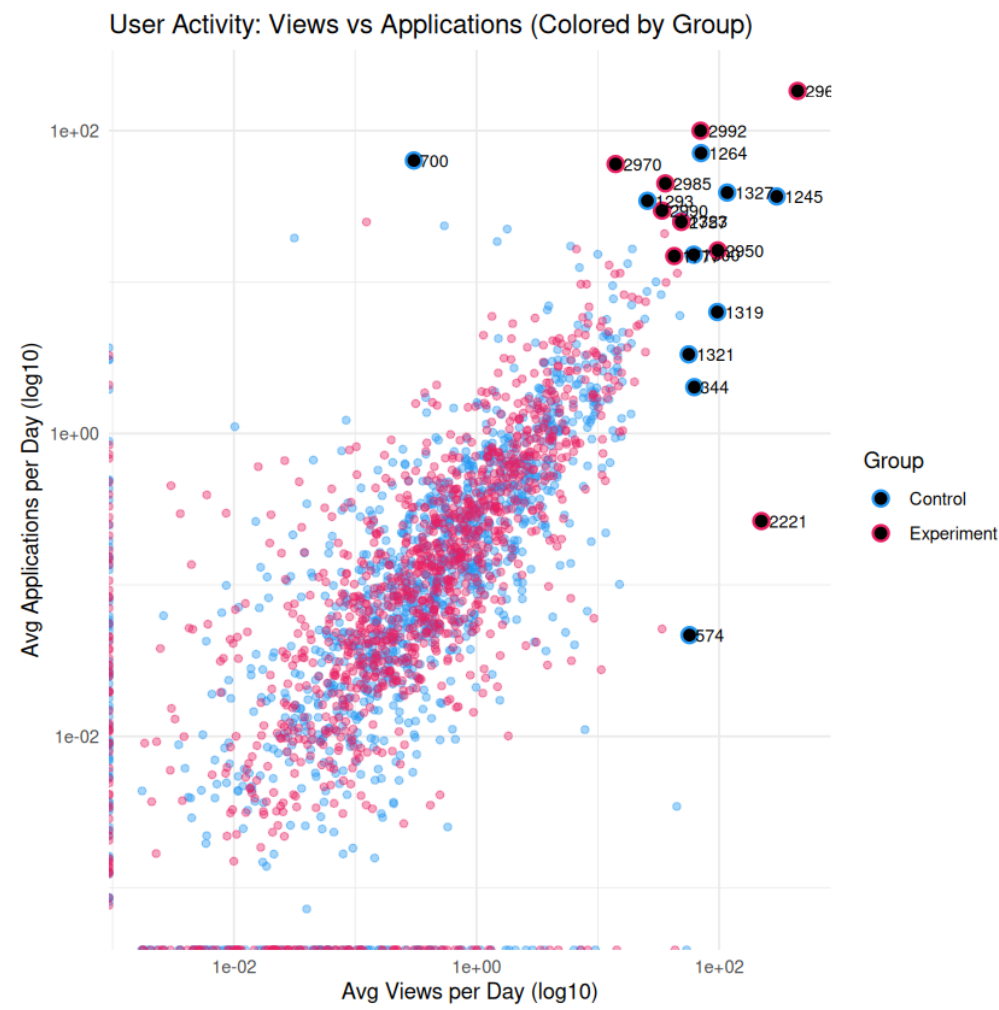
**Table 2: Difference‑in‑Differences (Control vs Experiment, Pre → Post)**

| **Metric** | **Term** | **β** | **SE** | **95 % CI** | **Sig** | **Interpretation** |
| --- | --- | --- | --- | --- | --- | --- |
| **Views / day** | Intercept (Ctrl‑Pre) | 2.83 | 0.33 | 2.19 – 3.47 | \*\*\* | baseline for control |
|  | Post (time) | 2.54 | 0.38 | 1.80 – 3.28 | \*\*\* | average pre→post jump shared by both groups |
|  | Experiment (group) | –0.04 | 0.46 | –0.94 – 0.86 | n.s. | exp vs ctrl before start |
|  | Post × Experiment | –0.10 | 0.53 | –1.14 – 0.94 | n.s. | DiD estimate (extra shift) |
| **Apps / day** | Intercept (Ctrl‑Pre) | 0.94 | 0.15 | 0.64 – 1.24 | \*\*\* | baseline for control |
|  | Post (time) | 1.18 | 0.18 | 0.83 – 1.52 | \*\*\* | shared pre→post jump |
|  | Experiment (group) | 0.12 | 0.21 | –0.30 – 0.54 | n.s. | exp vs ctrl before start |
|  | Post × Experiment | –0.17 | 0.25 | –0.65 – 0.32 | n.s. | DiD estimate |

Furthermore, we found no evidence in support of a boost in overall applications. Total job applications per user rose significantly in both groups after the intervention, but equally so. The treatment did not cause users to apply to more jobs in aggregate; treated users simply applied to a similar number of jobs as they would have without the notification. Importantly, we also did not observe a drop in applications – the intervention did not discourage users from applying to other jobs. A potential concern was that, after focusing on the “top” jobs, users might neglect other postings, but that did not happen. In summary, the intervention did *not* affect the volume of applications: it neither significantly increased nor decreased the number of jobs users applied to overall.

### **Overall Treatment Effect on Applications & Views without outliers**

Before our analysis, a preliminary inspection revealed several extreme observations (**Figure 5)** —users with exceptionally high daily views or applications. Such extremes, although relatively few, have the potential to distort mean comparisons and obscure the underlying treatment effect by inflating variance, thus reducing statistical clarity. To manage this issue systematically, we applied a standard statistical technique: the **1.5 × Interquartile Range (IQR)** method, which identifies and removes extreme outliers. Specifically, we removed user-job pairs whose daily views or applications exceeded 1.5 times the interquartile range (IQR) above the 75th percentile. This trimming approach is robust against skewed distributions, allowing us to generate more stable and interpretable estimates of the treatment effect.

****

**Figure 5: View/Applications outliers in control and experiment**

#### **Impact of Outlier Removal (Within-Group Analysis)**

After applying this outlier-filtering method, we performed regression analyses comparing pre-experiment (baseline) and post-experiment engagement within each experimental group. The results, summarized in **Table 3**, clearly indicate that both the control and experimental groups experienced substantial, statistically significant increases in activity between the pre- and post-periods. The clear and significant uplift in both groups confirms strong overall engagement trends during the experiment window. Notably, outlier removal reduced noise and variance, improving our ability to measure these shifts accurately.

**Table 3: Within‑group Pre → Post changes**

**(two regressions per group: one on views, one on applications)**

**(after trimming extreme outliers with the 1.5 × IQR rule)**

| **Metric** | **Group** | **Pre‑period mean (β₀)** | **Post shift Δ (β₁)** | **SE (Δ)** | **95 % CI** | **Sig** |
| --- | --- | --- | --- | --- | --- | --- |
| **Views / day** | **Control** | 0.79 | + 3.72 | 0.19 | 3.35 – 4.10 | \*\*\* |
| **Views / day** | **Experiment** | 0.77 | + 3.85 | 0.19 | 3.47 – 4.23 | \*\*\* |
| **Apps / day** | **Control** | 0.21 | + 1.55 | 0.08 | 1.39 – 1.71 | \*\*\* |
| **Apps / day** | **Experiment** | 0.25 | + 1.57 | 0.09 | 1.40 – 1.74 | \*\*\* |

#### **Subgroup Analyses: A Deeper Exploration**

Having identified robust increases within both experimental conditions post-treatment, we sought a deeper understanding of how specific user segments responded to the notification. We hypothesized that the notification’s effectiveness might vary depending on how competitively ranked the user was (e.g., top 4 vs. top 10) and the user’s confidence in their ranking chances. A good ranking in jobs with 30 or more applicants indicates stability of a high rank. Thus, we conducted further subgroup analyses focusing on these dimensions.

**Table 4-A: Views per day (during the experiment window)**

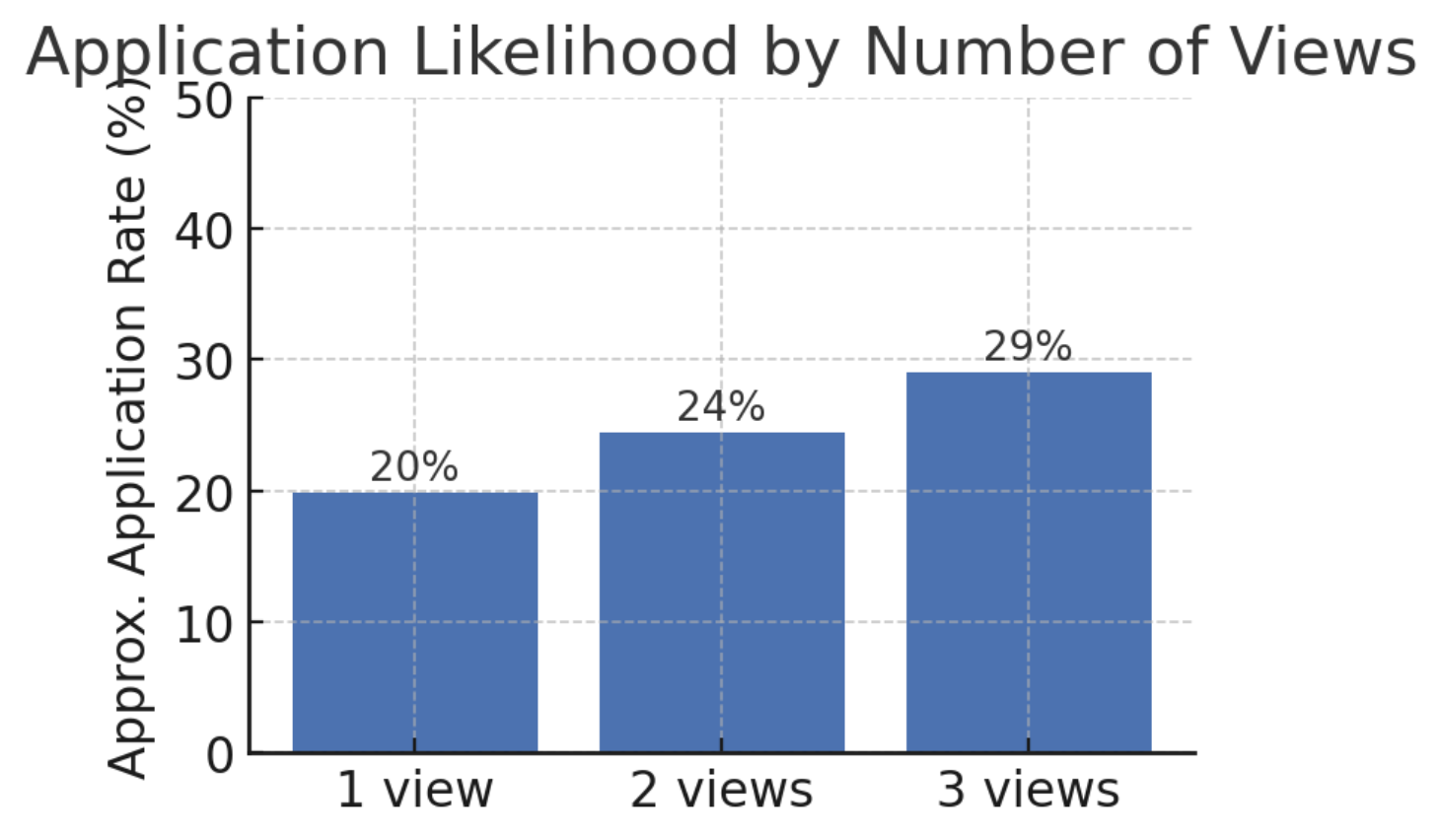
| **Sub‑group (rank / inbox)** | **Control mean** | **Experiment mean** | **Δ (Exp – Ctrl)** | **Std‑Err** | **95 % CI** | **p‑value** |
| --- | --- | --- | --- | --- | --- | --- |
| **≤ 4** | 0.323 | 0.387 | +0.064 | 0.023 | 0.020 – 0.109 | **0.005\*\*\*** |
| **≤ 10** | 0.326 | 0.355 | +0.029 | 0.016 | –0.001 – 0.060 | 0.060 |
| **≤ 4 & out\_of ≥ 30** | 0.261 | 0.360 | +0.099 | 0.034 | 0.033 – 0.166 | **0.004\*\*\*** |
| **≤ 10 & out\_of ≥ 30** | 0.305 | 0.344 | +0.039 | 0.019 | 0.001 – 0.076 | **0.043** |
| **All jobs** | 0.306 | 0.327 | +0.021 | 0.008 | 0.005 – 0.038 | **0.012\*\*** |

These subgroup analyses reveal a clear pattern: the notification’s effect on views intensifies as the ranking becomes more believable, with more applicants in the box, and the user's ranking is still high. For instance, in scenarios **(rank ≤ 4 and ≥ 30 applicants)**, notified users viewed significantly more jobs **(+0.099 views/day, p = 0.004)**. Thus, while the notification generally increases interest, it does so most powerfully when users perceive the ranking as more believable with already enough candidates in the inbox.

**Table 4-B: Applications per day (during the experiment window)**

| **Sub‑group (rank / inbox)** | **Control mean** | **Experiment mean** | **Δ (Exp – Ctrl)** | **Std‑Err** | **95 % CI** | **p‑value** |
| --- | --- | --- | --- | --- | --- | --- |
| **≤ 4** | 0.125 | 0.119 | –0.006 | 0.014 | –0.032 – 0.021 | 0.666 |
| **≤ 10** | 0.153 | 0.156 | +0.003 | 0.009 | –0.014 – 0.021 | 0.706 |
| **≤ 4 & out\_of ≥ 30** | 0.188 | 0.189 | +0.001 | 0.017 | –0.033 – 0.035 | 0.957 |
| **≤ 10 & out\_of ≥ 30** | 0.180 | 0.192 | +0.012 | 0.010 | –0.008 – 0.032 | 0.244 |
| **All jobs** | 0.175 | 0.179 | +0.004 | 0.005 | –0.005 – 0.013 | 0.400 |

However, this increased engagement did not fully translate into proportional increases in applications for those jobs. In all the subsets analyzed, the **conversion from viewing to applying remained relatively low**. Treated users were only slightly, and not significantly, more likely to apply to the jobs they viewed than control users. For instance, even when a user was ranked less than or equal to the 10th position out of a large pool (which the product team assumed would be an enticing scenario), viewing rates went up with the treatment but application rates did not show a reliable rise. This suggests an **engagement-action gap**: users would click and browse the recommended jobs out of curiosity or interest, but many ultimately decided not to apply. Several plausible user behaviors could explain this. Users may be **more selective when deciding to apply** – even if they know they are a top candidate, they will only apply if the job genuinely appeals to them, considering factors such as role, salary, location, etc. Being highly ranked doesn’t guarantee the job is desirable; it only indicates the user’s fit relative to other applicants. Another insight is that **job attractiveness moderates behavior**. If the recommended job is not attractive to the user, the top-candidate label alone won’t convince them to apply. Thus, the treatment primarily affected the *top of the funnel* (views) but not the *bottom of the funnel* (conversions to applications). One encouraging sign is that repeated exposure appeared to help – users who saw the same “top job” multiple times (or on multiple visits) were more likely to apply eventually, suggesting that **repetition can increase engagement** to the point of action **(Figure 6)**. This assumption, although plausible, may need further validation, as repeated views could also indicate an interest that prompted the user to view the job multiple times, or possibly storing it and viewing it before applying. But within our experiment window, single-session conversions remained limited.

****

**Figure 6: Applications/views**

### **Subgroup and Covariate Analysis**

We also examined user characteristics to determine if the notification was more effective for specific types of job seekers. Adding **matching covariates** (such as whether the user met the job’s experience requirement or location match) in the regression did not substantially change the estimated treatment effect – it remained near zero. Those models revealed some interesting relationships: for example, application likelihood was strongly influenced by **location** **match and experience;** users were far more likely to apply if the job was in their country and if they met the required experience range. This underscores that a good fit drives applications, which is intuitive. It also means that our treatment often highlighted jobs that were a good fit (by design), so those jobs were inherently more likely to receive applications from both groups.

Another observation was counterintuitive: some **less-qualified candidates tended to apply aggressively**, even though they were not top candidates for many jobs. In other words, a user might be ranked highly for one job, making them eligible for our experiment, but actually be underqualified in some aspect. We saw cases where such users still applied widely, even to jobs for which they weren’t top-ranked. This suggests that **not all users behave “rationally” based on fit feedback – some will apply it** broadly regardless of the input. This could dilute the effect of our intervention for that segment, since those users would apply a lot with or without extra prompts. Overall, incorporating covariates improved the precision of our estimates (reducing residual noise) but confirmed that **no hidden treatment effect** was lurking after accounting for user and job attributes. The coefficient for the treatment indicator remained statistically nonsignificant in all enhanced models.

In summary, **the experiment’s core finding is null**: the top-candidate notification did *not* produce a statistically significant increase in job applications. The treatment did increase job views (engagement with recommended jobs), showing that users noticed and interacted with the feature. However, this did not translate into a measurable uptick in application submissions. The result is consistent across various subsets and remains so after controlling for multiple factors. The evidence suggests that while users were interested in the information (“You are a top candidate”), it alone was not a strong enough motivator to change their final application decisions in the short term.

**Table 5-B: Covariate‑adjusted OLS results ( post‑period, all jobs subset) - Views**

| **Term** | **β** | **SE** | **95 % CI** | **Sig.** |
| --- | --- | --- | --- | --- |
| **Intercept** | 1.227 | 0.094 | 1.042 – 1.412 | \*\*\* |
| **group = Experiment** | 0.0788 | 0.0278 | 0.024 – 0.133 | \*\* |
| **seq** | 0.1926 | 0.0122 | 0.169 – 0.216 | \*\*\* |
| **career\_level not matching** | –0.172 | 0.044 | –0.259 – –0.086 | \*\*\* |
| **experience not matching** | –0.0855 | 0.034 | –0.152 – –0.019 | \* |
| **age not matching** | 0.1087 | 0.080 | –0.049 – 0.267 | n.s. |
| **location not matching** | –0.613 | 0.070 | –0.751 – –0.475 | \*\*\* |
| **education not matching** | –0.060 | 0.035 | –0.129 – 0.009 | n.s. |
| **gender not matching** | –0.179 | 0.043 | –0.262 – –0.095 | \*\*\* |
| **nationality not matching** | 0.007 | 0.037 | –0.066 – 0.081 | n.s. |

**Table 5-B: Covariate‑adjusted OLS results ( post‑period, all jobs subset) - Applications**

| **Term** | **β** | **SE** | **95 % CI** | **Sig.** |
| --- | --- | --- | --- | --- |
| **Intercept** | 0.427 | 0.043 | 0.342 – 0.512 | \*\*\* |
| **group = Experiment** | 0.0191 | 0.0132 | –0.0067 – 0.0450 | n.s. |
| **seq** | 0.0475 | 0.0051 | 0.0376 – 0.0574 | \*\*\* |
| **career\_level not matching** | –0.030 | 0.0296 | –0.088 – 0.028 | n.s. |
| **experience not matching** | –0.0521 | 0.0167 | –0.0848 – –0.0194 | \*\* |
| **age not matching** | –0.062 | 0.035 | –0.131 – 0.007 | n.s. |
| **location not matching** | –0.166 | 0.0270 | –0.219 – –0.113 | \*\*\* |
| **education not matching** | –0.017 | 0.019 | –0.054 – 0.020 | n.s. |
| **gender not matching** | –0.0228 | 0.0173 | –0.0566 – 0.0111 | n.s. |
| **nationality not matching** | –0.0174 | 0.0208 | –0.0580 – 0.0233 | n.s. |

## **Discussion and Interpretation**

**Why didn’t applications increase?** The lack of a treatment effect on applications prompts essential considerations. One interpretation is that **being informed of one’s high rank is not, by itself, a sufficient incentive to apply**. Users might already have an internal sense of which jobs they are a good fit for, based on the job description, requirements, and their qualifications. The notification essentially confirmed something they might suspect (or, if it was a surprise, they still weigh the job’s merits). If a job listing is attractive and the user feels qualified, they likely would apply regardless of seeing a “top candidate” badge. Conversely, if the job isn’t appealing, knowing they rank highly among applicants doesn’t overcome that fundamental lack of interest. Thus, the intervention may have been addressing the wrong bottleneck – the **limiting factor in application submission might be job attractiveness or personal preference, not just confidence or awareness of fit**. Our data supports this: treated users *saw* the jobs, indicating awareness was raised, but many still chose not to apply.

Additionally, **user behavior nuances** might have diluted the measurable impact. Some users in the treatment group may not have visited the site frequently or may have ignored the notifications. However, since compliance was automatic, the primary dilution would come from users who simply didn’t have many opportunities or didn’t need the suggestion. Notably, we restricted to users with at least one top-candidate job; a byproduct is that these users likely already had decent application activity (they had applied to trigger the treatment). This group might be relatively savvy or motivated job seekers. It could be that less-active users (who were not in our sample because they hadn’t applied to anything yet) might have been more dramatically influenced by a feature like this – e.g., if someone is discouraged and not applying at all, telling them “here’s a job where you’re a top candidate” could spur them into action. Our experiment couldn’t measure that segment, since by design, we needed users to apply to something to receive the treatment. In a sense, we **tested the intervention on already-active applicants**, not on completely passive seekers. So, one interpretation is that the feature might have more impact as a *re-engagement tool* for low-activity users, which was outside our test scope.

**Design implications:** Despite the neutral outcome on the primary metric, there are useful insights for designing future product interventions and experiments. The fact that **users did engage with the recommendations** (increased views) indicates that the concept of surfacing “high-fit” jobs is attractive. The challenge is converting that interest into action. Product teams might consider augmenting the top-candidate notification with additional persuasive elements – for example, including a brief highlight of *why* the user is a top candidate (skills match or experience) to increase their confidence, or providing an incentive (“High match! Your chances are great to be viewed by the employer – apply now!”). In other words, **a richer messaging strategy might be needed to push users from browsing to applying**. This aligns with one of our brainstorming ideas: combine the ranking feedback with supportive messaging to counter potential hesitation. For instance, if a user is rank 5 out of 10, they might think the competition is still large; a message emphasizing their strong qualifications or encouraging them not to miss the opportunity could help. Future A/B tests could compare the plain rank notification versus an enhanced motivational message.

From an experimental design perspective, this study underscores the importance of **adequate sample size and timing**. If we suspect external trends, using a **shorter experiment window or a more stable period** could help isolate the treatment effect. Alternatively, employing a **blocking or stratified randomization** by time or user activity might help account for temporal spikes. We could also measure outcomes over a longer follow-up to see if the treatment had delayed effects (perhaps users flagged those jobs and applied later). In this experiment, we primarily captured immediate and short-term behavior; it’s possible that some users returned days later to apply after thinking it over, which might not fully show in our analysis if the window was limited.

It’s also worth noting that in a live product rollout, **users who never see a top-candidate notification (because they are never in the top 10 for any job) might experience the platform differently**. While not directly in our experiment, this raises a broader design question: should we provide some feedback to those users as well (perhaps a different kind of guidance to deter them)? Otherwise, the feature only engages already-strong candidates and leaves others without any new input. Ensuring the system benefits a wide range of users would be important for overall platform health (even if it means advising some users to improve their profile or search in different categories, rather than showing “top candidate” status).

**Business interpretation:** For Bayt.com, the key takeaway is that simply informing candidates of their relative rank did not boost application numbers in the short term. This means the feature, as tested, may not directly increase site KPI’s like total applications. However, it did increase user engagement with job content (views) and did not have any negative effects, which suggests it provided value (users were interested in the information). It might improve user satisfaction – we can speculate that job seekers appreciated knowing where they stand, even if it didn’t make them apply more. That kind of qualitative benefit wasn’t measured here but could be gleaned from follow-up surveys. Also, even without increasing the number of applications, there could be a **quality benefit**: if users slightly shift their applications toward jobs where they are a better fit (and perhaps away from long-shot applications), the hiring outcomes might improve, with a higher success rate and less wasted effort. Our experiment did not track ultimate outcomes, such as interview callbacks or job offers, but it’s plausible that focusing candidates on roles where they rank highly could yield better matches. For future analysis, it would be interesting to see if the employers for those jobs notice any difference in candidate quality or if time-to-hire improves due to more qualified applicants applying.

**Limitations:** We have already discussed power and timing as limitations. Another limitation was that we measured only on-site behavior. Some users may have seen the notification and chosen to apply via other means (e.g., if they saw the company name and used offline or reached out to the hiring manager through social media) – although unlikely, this would evade our tracking. We assume that it didn’t occur at scale. Additionally, our observation period was relatively short, covering the immediate post-intervention weeks. It’s possible the treatment’s impact could manifest later; for example, a user could remember they were top-ranked and revisit the job after a delay. A longer-term measurement could capture if there were any persistent effect on engagement or applications (our data suggests not, but we can’t be certain beyond the observation window).

Lastly, **generalizability** is a consideration. This experiment was conducted on a specific platform and with a particular user base. The behavior of job seekers may differ on other platforms or in different regions. For instance, in some markets, candidates might be more risk-averse, and the reassurance of being a top candidate could have a bigger impact. Or the opposite: perhaps Bayt’s users are highly active and savvy (many applications per day), leaving little room for this nudge to change them, whereas on a platform with more passive seekers, a nudge could have more effect. In future experiments, segmenting by user type (active vs passive job seekers) could be informative.

## **Conclusion**

This randomized experiment provided a rigorous test of a “top candidate notification” feature in a real online job marketplace. The results indicate that **merely notifying users that they are** **top-ranked among applicants did not significantly increase their likelihood of applying to** **those jobs**. Users did respond to the notification by viewing more suggested jobs, but that interest largely stopped short of converting into additional applications. The findings highlight a gap between **engagement and action** – awareness of fit alone may not overcome other factors in the decision to apply. From a causal inference standpoint, our study benefited from clear identification (random assignment and perfect compliance) and thus provides credible evidence that the observed null effect is indeed the outcome of the intervention (given the context and sample).

For practitioners, this suggests that while transparency features like rank feedback can enhance user experience, they might not be a silver bullet for boosting participation. Improvements or complementary features, such as more personalized encouragement or broader eligibility, may be needed to achieve the desired increase in applications. It’s also possible that the value of this feature lies in quality and satisfaction rather than the quantity of applications – something a single experiment focused on counts cannot fully capture. In terms of product development, the concept of guiding job seekers toward roles where they have better chances remains sound (it aligns with user desires for clarity). Still, it should be implemented in a way that more directly motivates people to take action.

In conclusion, our experiment did not find evidence of a causal impact on short-term application behavior from the top candidate notifications. However, it yielded important insights into user behavior. Candidates are willing to engage with tailored recommendations, but additional steps are required to translate that into outcomes. Future experiments can build on these insights – for example, by testing enhanced messaging, different eligibility criteria, or examining longer-term effects such as the quality of matches. Ultimately, combining behavioral data with thoughtful experimental design will help refine features that effectively support job seekers in making better applications without overwhelming them. The findings from this study contribute to a broader understanding of how feedback interventions can shape (or, in this case, *not* shape) user behavior on online platforms.