Hybrid, Remote, or On–site? Causal Evidence from a Multi-Firm Rotating-Exposure Study in Greece

Christina Kovatsi*

Areimanio Tutoring School and Independent Researcher, 126 Lavriou Ave., Glika Nera 153 54, Greece; e–mail: kovatsi.christina@gmail.com

*Corresponding Author: kovatsi.christina@gmail.com

DOI: https://doi.org/10.30209/IJMO.202503.018

Submitted: Oct. 14, 2025 Accepted: Nov. 26, 2025

ABSTRACT

Organizations need credible, decision-ready evidence on when flexible work arrangements improve outcomes. We estimate the causal effects of hybrid and fully remote schedules, relative to on-site, on employee productivity, job satisfaction, and Work–Life Climate (WLC). A longitudinal, rotating-exposure design was implemented across three multinational firms in Greece (36 teams, 537 employees), with intact teams cycling through on-site, hybrid, and remote conditions. Administrative performance indicators and validated surveys were collected alongside qualitative diaries and focus groups. Identification used difference-in-differences models with event-study diagnostics. Hybrid produced modest gains in objective productivity (~+0.07 standard deviations [SD]), larger improvements in subjective productivity (~+0.13 SD) and job satisfaction (~+0.12 SD), and a better WLC (~−0.09 SD). Fully remote averaged near zero on productivity. Effects were largely mediated by increased autonomy and improved task–technology fit, attenuated for high-complexity roles, and amplified under stronger managerial support. Results were robust across alternative estimators and inference adjustments. Practically, flexibility performs best as a designed system: set a predictable in-person cadence for interdependent work, institutionalize supportive management routines, and integrate digital tools to sustain coordination and well-being.

Keywords: Hybrid work, Remote work, Productivity, Job satisfaction, Work–life climate, Difference–in–differences, Task–technology fit

1. Introduction

The rise of Flexible Work Arrangements (FWAs) has transformed how organizations coordinate effort, allocate autonomy, and maintain cohesion. After the global, involuntary experiment in remote work, many employers moved toward hybrid schedules that combine periodic in–person collaboration with at–home work, while others retained fully remote or reverted to on–site patterns. The empirical record, however, remains uneven across outcomes, contexts, and identification strategies. Recent large–scale field evidence indicates that hybrid schedules can improve retention

and satisfaction without harming performance [1]. In contrast, fully remote work has been found to increase network siloing and reduce cross–team bridging ties, a potential brake on productivity when interdependence is high [2]. Complementary studies show that work–from–anywhere can increase output when tasks are modular, and performance management is digitized [3]. Macro evidence documents the durable rise of hybrid WFH, driven by worker preferences and firm learning rather than a temporary shock [4]. These findings suggest that the "effect of flexibility" is not a constant, but a function of task structure, coordination design, and managerial scaffolding.

This paper investigates the causal impact of three flexibility modalities (on–site, hybrid, and fully remote) on employee productivity (objective and subjective), job satisfaction, and WLC. We study three multinational firms operating in Greece and leverage a rotation–based design in which intact teams transition through pre–specified exposure blocks, enabling modern Difference–in–Differences (DiD) identification with event–study diagnostics that address staggered timing and treatment heterogeneity [5], [6]. Our measurement strategy pairs administrative KPIs (e.g., project throughput, sales conversions, DevOps/DORA indicators) with validated survey instruments. For productivity perceptions, we use the WHO HPQ and the Individual Work Performance Questionnaire (IWPQ) [7], [8]. For job satisfaction, the Job Descriptive Index / Job–in–General (JDI/JIG) [9], [10], with Greek validation [11]. For WLB, we apply a Work–Life Climate scale that emphasizes boundary behaviors rather than moods [12], [13]. To illuminate mechanisms, we assess autonomy, communication structure/frequency, and task–technology fit (TTF), a classic lens that argues that performance hinges on alignment between task requirements and tool capabilities [14]. Qualitative diaries and focus groups provide process evidence that we formally integrate with the quantitative results using joint displays [15], [16].

Three tensions in the literature motivate our approach. First, studies often trade off internal validity for external validity. Single–firm experiments or quasi–experiments may pin down causal effects but raise questions about generalizability across roles and technologies; cross–sectional surveys generalize more easily but struggle with selection and confounding. Our multi–firm, rotating–exposure design directly addresses this trade–off, combining credible identification with organizational diversity (role families spanning engineering, sales, and operations). Second, prior work typically emphasizes either hard performance metrics or employee experience. We measure both, recognizing that subjective productivity and satisfaction may respond earlier or more strongly than objective output, especially when commute relief and schedule control reduce time fragmentation [7], [8], [9], [10], [12], [13]. Third, much empirical debate collapses mechanisms into a single "remote/hybrid effect." We decompose outcomes through mediators (autonomy, communication, TTF) and boundary conditions (role complexity, managerial support), connecting micro–mechanisms to macro effects [17], [18], [19], [14].

Beyond these empirical tensions, there is also a theoretical gap. Self–determination theory and work–design meta–analyses predict that greater autonomy should improve motivation, satisfaction, and often performance, particularly when employees can schedule and sequence tasks more freely [20], [21]. Task–technology fit research predicts that once tools are well aligned with task demands,

performance differences across locations should shrink [14]. Coordination perspectives instead emphasize that when tasks are complex and tightly coupled, the architecture of coordination, including co-location, routines, and shared artefacts, becomes a dominant constraint [22], [23]. Existing studies of flexible work arrangements typically invoke one of these perspectives at a time and often treat "remote" or "hybrid" as black–box treatments. Few provide joint tests of how autonomy, communication structure, and digital fit interact with role complexity and managerial support, or identify conditions under which the theories yield conflicting predictions. Our study explicitly positions FWAs as a moderated–mediation system that synthesizes and partially extends these frameworks by estimating when the autonomy and fit channels are strong enough to overcome the coordination demands implied by complex, interdependent work.

Our research questions are therefore: (1) *What are the intent–to–treat effects of hybrid and fully remote work (vs. on–site) on objective productivity, subjective productivity, job satisfaction, and WLC?* (2) *To what extent are these effects mediated by autonomy, communication, and TTF?* (3*) How do effects vary with role complexity and managerial support?* (4) *Are the results robust under alternative estimators and exposure definitions, and do event–study pre–trends support parallel–trends assumptions?* These questions organize the paper's design and analysis.

Our contributions are fourfold. First, we provide credible, multi–firm causal evidence on FWAs using a staggered rotation and modern DiD estimators (Sun–Abraham; Callaway–Sant'Anna) that explicitly handle treatment heterogeneity and timing [5], [24]. Second, we deliver a multi–method outcome battery. Objective, administrative KPIs and collaboration traces alongside HPQ/IWPQ, JDI/JIG, and Work–Life Climate [7], [8], [9], [10], [12], [13], which allows us to compare "hard" and "perceived" performance instead of treating one as a proxy for the other. Third, we integrate qualitative mechanisms with quantitative dynamics via joint displays, aligning relative–time ATT_k patterns with theme intensity to reduce post–hoc narrative risk and to show when and why effects emerge [15], [16]. Fourth, we surface boundary conditions that help reconcile mixed findings in the literature. Hybrid tends to outperform on affective and perceived outcomes and shows small gains on objective metrics, provided that interdependent work benefits from periodic co–location and that managerial support sets clear priorities and boundary–respecting norms [19], [18], [14], [1], [2], [3], [4]. Where role complexity is high and coordination guardrails are thin, fully remote effects tend to be neutral on average, consistent with network siloing under distance [2].

The remainder of the paper proceeds as follows. Section 2 reviews the literature on FWAs, productivity, and well–being measurement, and mechanism theories, alongside recent field evidence on hybrid/remote outcomes. Section 3 develops the conceptual model and hypotheses. Section 4 details the setting, rotation protocol, measures, and identification strategy. Section 5 reports descriptive statistics and main DiD estimates, validates pre–trends via event–studies, integrates mechanisms using mixed–methods, and examines boundary conditions and robustness. Section 6 concludes with theoretical and managerial implications, situating our results alongside recent randomized and quasi–experimental studies.

2. Literature Review

2.1 Flexible Work Arrangements

FWAs in knowledge–intensive firms are commonly organized as entirely on–site (all paid days at the employer's premises), fully remote (all paid days off–premises), or hybrid (a stable mix of on–site and off–premises days within a week or cycle). In EU statistics and labor research, "telework" denotes work performed away from the employer's premises using information and communication technologies. "Regular telework" corresponds to what most organizations call hybrid (recurring off–premises days), whereas "occasional telework" captures sporadic remote days [25], [26]. These distinctions matter empirically because incidence, worker composition, and outcome patterns differ across the three modalities.

Across the EU–27 in 2024, about 3% of employees engaged in full–time telework, ~9% in regular/hybrid telework, and ~16% in occasional telework [25]. Telework is concentrated among managers (≈60%), professionals (≈58%), and technicians (≈40%), with notably higher rates in financial services and education, reflecting the digitalizability of tasks and occupational autonomy [25]. These descriptive patterns align with cross–country evidence from the Global Survey of Working Arrangements (G–SWA) and related analyses, which document wide heterogeneity in WFH intensity and a persistent shift toward hybrids among college–educated workers in advanced economies [27], [28]. Explanations for cross–national variation include cultural individualism, lockdown stringency, population–weighted density, sectoral mix, and incomes [28].

In Greece, telework uptake remains below the EU average. In 2023, 1.9% of Greek employees "usually" worked from home and 5.5% "sometimes" did so, compared with 8.9% and 13.3% respectively in the EU–27 [29]. These figures, drawn from official Eurostat labor force statistics, underscore both the headroom for expansion and the importance of Greece's sectoral composition (e.g., tourism and retail) for feasible WFH penetration.

2.1.1 Prior findings on productivity

Pre–pandemic, the best–identified evidence came from a randomized experiment at a Chinese call center (CTrip), which found a 13% performance increase under work–from–home. Roughly 9 percentage points from more minutes worked (fewer breaks/absence) and ~4 points from higher per–minute calls in quieter home environments; job satisfaction improved and quit rates fell [30]. In contrast, during the pandemic's abrupt, fully remote shift in a large Asian IT services firm, a high–frequency personnel/analytics panel showed 8–19% lower productivity (output per hour), driven by longer hours, more coordination time, and fewer uninterrupted blocks for deep work [31]. These studies imply that productivity effects are context– and task–dependent. Individually executed, measurable tasks can benefit from WFH, whereas collaborative, interdependent work can experience coordination frictions at full distance.

The most policy–relevant causal evidence post–pandemic is a six–month randomized controlled trial of hybrid scheduling (two WFH days/week) among 1,612 graduate employees in engineering, marketing, and finance. The study finds no detectable effect on performance reviews or promotions over two years, no impact on software engineers' code output, higher job satisfaction, and a one–third reduction in attrition, with larger retention gains among non–managers, women, and longer

commuters [1]. Managers' pre–trial priors (−2.6% perceived productivity effect) updated toward neutrality/positivity after exposure to hybrid practices [1]. These findings suggest that hybrid arrangements can maintain measured performance while improving retention and satisfaction in team–based creative work.

At the macro/organizational level, syntheses by international bodies report broadly positive short–term assessments of telework for self–rated productivity and well–being, while cautioning that evidence on long–run productivity effects remains limited and potentially non–linear. Benefits rise with moderate telework but may erode at high intensities due to coordination and learning costs [27], [32]. This non–linearity is consistent with micro–evidence. Fully remote arrangements can raise communication costs for complex collaboration [31], whereas hybrid schedules can preserve co–located coordination for problem–solving, onboarding, and tacit knowledge transfer. Recent work on proximity to coworkers shows that co–location increases mentoring and feedback, especially for junior engineers and women, improving skill accumulation even when short–run output might not increase, highlighting a training–throughput trade–off [33].

2.1.2 Prior findings on job satisfaction and work–life balance

Decades of organizational research, including pre–pandemic meta–analyses, find small but positive average effects of telecommuting on job satisfaction and affect, mediated by greater autonomy and reduced commuting; drawbacks include risks of social isolation and boundary blurring when intensity is high [34], [35]. Post–pandemic reviews converge on a similar pattern. Many workers enjoy remote/hybrid work, but well–being varies with managerial support, communication norms, and home–office ergonomics; isolation and blurred boundaries are the most common stressors [36]. Recent EU–27 data confirm that regular (hybrid) and occasional teleworkers are more likely to work during free time or be contacted outside hours, indicating a need for guardrails when scaling flexibility [25]. In the hybrid RCT noted above, job satisfaction increased without harming performance, and employees cited commuting–time savings and schedule control as key contributors [1].

2.2 Measurement of Productivity and Job Satisfaction / Work–Life Balance

In this study, "productivity" is operationalized using both objective and subjective indicators to balance precision with ecological validity. Objective indicators are sourced from firms' performance-measurement systems and mapped to role families. Project throughput and on–time completion for project roles, conversion rates, and quota attainment for sales, and software delivery throughput–stability metrics (e.g., deployment frequency, lead time, change fail rate, time to restore) for engineering teams. The latter are widely adopted, empirically grounded measures of software delivery performance and its link to organizational outcomes [37]. More generally, key performance indicators (KPIs) are defined within performance–management systems to reflect critical success factors rather than mere activity counts [38].

Subjective productivity is captured via brief, repeated self–reports calibrated to established instruments. The WHO Health and Work Performance Questionnaire (HPQ) provides validated self–ratings of absolute and relative performance, absenteeism, and presenteeism, with published evidence

for reliability, criterion validity against employer records, and sensitivity to change across occupations [7], [39]. Complementarily, the IWPQ provides a generic behavioral measure encompassing task performance, contextual performance, and counterproductive work behavior, with acceptable construct validity across diverse worker samples [8]. These scales reduce idiosyncratic wording variance and facilitate cross–role comparisons when objective output units are incommensurable.

"Job satisfaction" is measured with the Job Descriptive Index (JDI) and the Job–in–General scale, which separately assess satisfaction with work, pay, promotion, supervision, and coworkers, as well as global job satisfaction. The JDI is extensively validated and periodically updated, with strong evidence for construct validity and international use; a recent Greek adaptation supports local psychometric adequacy, aiding interpretability in the present context [9], [10], [11]. "Work–life balance" is assessed using the Work–Life Climate Scale, which emphasizes observable boundary behaviors (e.g., skipping meals, after–hours work), shows solid psychometric properties, and correlates with safety climate and burnout in large healthcare samples, useful when distinguishing behavioral norms from affective states [12], [13].

To mitigate inflation from common–method variance when self–reports co–move, subjective measures are scheduled separately from satisfaction/climate batteries, combined with objective KPIs at the analysis stage, and accompanied by recommended procedural and statistical remedies from the measurement literature [40]. This multimethod specification allows triangulation. Objective KPIs anchor role–specific outputs, while validated self–reports capture perceived efficiency and constraints that firm systems may not record.

2.3 Mechanisms

Three mechanisms recur across organizational research as proximate drivers of outcomes under flexible work. Autonomy increases self-determined motivation and facilitates self-regulation, which in turn supports performance and positive affect. Meta–analytic evidence links autonomy to higher satisfaction and task performance, while self–determination theory explains the pathway through basic need satisfaction and the internalization of work goals [20], [21]. Leader behaviors that *support* autonomy (e.g., providing choice, acknowledging perspectives) amplify these effects and are themselves positively associated with employee motivation and well–being, strengthening the autonomy–outcome channel in dispersed settings [41].

Communication frequency and structure shape how information travels when co–presence is reduced. Large–scale causal evidence from a firm–wide remote shift shows that collaboration networks become more static and siloed, with fewer bridging ties across groups. These patterns can dampen information recombination even as asynchronous traffic rises [2]. Laboratory and field studies further indicate that videoconferencing narrows attentional scope, reducing idea generation relative to in–person interaction, cautioning against assuming medium–independence in creative work [42]. These findings imply that communication cadence and media choice are not neutral. They alter network topology and cognitive breadth in ways that matter under hybrid or remote designs.

Digital tool efficacy conditions whether autonomy and communication translate into output.

Task–technology fit theory predicts that tools improve performance when their functionality matches task demands. Conversely, misfit introduces frictions (e.g., rework, coordination lag) that offset the gains from flexibility [14]. A sociomaterial perspective adds that technologies afford and constrain action repertoires, so configuration (integrations, defaults, visibility) co–evolves with routines and affects effectiveness in distributed teams [43].

Two boundary conditions are especially salient. Role complexity/interdependence increases coordination requirements; meta–analytic and review evidence show that higher task interdependence shifts performance dependence toward rich coordination mechanisms (shared understanding, temporal routines), making outcomes more sensitive to how hybrid or remote collaboration is orchestrated [22], [23]. Managerial support moderates' strain and sustains engagement. Perceived organizational support exhibits robust positive associations with satisfaction and commitment, while family–supportive supervisor behaviors buffer work–nonwork conflict, both of which are relevant for maintaining well–being under flexible schedules [44], [19]. Overall, self-determination theory, task–technology fit, and coordination perspectives make different conditional predictions about flexible work. Autonomy and supportive supervision suggest that greater discretion generally improves motivation, satisfaction, and performance [20], [21], [41], [44], [19]. Task–technology fit points to the quality of the collaboration stack as a bottleneck or amplifier of these gains [14], [43]. Coordination theory highlights that when tasks are highly interdependent, the costs of misaligned communication patterns or fragmented tools can outweigh the benefits of autonomy [22], [23]. What is missing in the empirical literature is an integrated, testable formulation that quantifies these channels and their boundary conditions within a single design. The present study addresses this gap by using the rotating–exposure setting to estimate a conditional process in which flexibility shapes autonomy, communication structure, and digital fit, and in which role complexity and managerial support determine whether these mechanisms translate into productivity and well–being outcomes.

2.4 Gaps and Innovation

Empirical work on FWAs has grown quickly, but much of it remains single-firm, single-modality, or cross-sectional, which limits both identification and transportability of effects. At the theoretical level, most studies adopt either a self-determination/work-design lens, a task–technology fit perspective, or a coordination/interdependence frame, with little effort to specify how these accounts intersect or when each should dominate [20], [21], [14], [22], [23]. Our design addresses both empirical and theoretical gaps on four fronts. First, rotating exposure assigns employees to hybrid, fully remote, and on-site conditions in sequence, enabling within-person contrasts while preserving organizational feasibility. Methodologically, this draws on stepped-wedge and crossover logics, gradual rollout across units combined with within-unit/within-person comparisons, which improve internal validity when simultaneous randomization is impractical, provided time trends and carryover are handled explicitly [45], [46], [47].

Second, a multi–firm setting strengthens external validity by allowing for the observation of effects across heterogeneous contexts (industries, roles, managerial practices). This responds to well–known cautions that internally valid estimates often travel poorly unless designs actively sample

variation and theorize scope conditions [48], [49].

Third, we adopt modern DiD estimators that account for staggered timing and heterogeneous effects. Two–way fixed–effects (TWFE) estimators can mix comparisons with problematic or negative weights when timing varies; decomposition and diagnostics clarify these weights and their implications [6]. To avoid such distortions, we estimate group–time average treatment effects and event–study dynamics using procedures that remain valid under treatment–effect heterogeneity and variation in adoption timing [5], [24]. This combination (rotation plus multi–period DiD) tightens identification while transparently separating time effects, exposure intensity, and role–level covariates.

Fourth, we use a multi–method strategy that integrates objective KPIs with validated self–reports (Section 2.2) and qualitative mediators from diaries and focus groups (Section 3). Integration will follow established mixed–methods practices. Joint displays align quantitative effects (e.g., DiD estimates by role complexity) with co–occurring qualitative mechanisms (e.g., shifts in autonomy or communication patterns), supporting mechanism–focused inference rather than parallel narratives [15], [16]. This integration does not substitute for statistical mediation; instead, it triangulates pathways and boundary conditions to explain why the same policy can produce different productivity–satisfaction profiles across teams. Conceptually, this design also allows us to move beyond descriptive catalogues of mechanisms toward an integrated test of how work–design, task–technology fit, and coordination perspectives jointly perform in a flexible work setting. By estimating mediated and moderated effects within the same difference–in–differences and event–study structure and by anchoring them in qualitative process evidence, we can assess whether autonomy and perceived fit are sufficient to offset coordination costs at different levels of interdependence and support. Therefore, the study both synthesizes existing constructs and sharpens their scope conditions, identifying where self-determination and TTF predictions hold, where coordination demands constrain them, and where the three perspectives converge.

3. Theoretical Framework and Hypotheses

3.1 Conceptual Model

Figure 1 posits a causal chain from the flexibility condition (on-site, hybrid, fully remote) to three outcome domains, productivity, job satisfaction, and work–life climate, operating through three mechanisms and two boundary conditions. First, autonomy. Flexible schedules increase discretion over effort allocation and timing; under self-determination theory, greater autonomy supports internalized motivation and task persistence, predicting higher performance and satisfaction when discretion can be exercised without undermining coordination [21]. Second, communication frequency/structure. Shifts to distance work rewire collaboration networks, reducing bridging ties and making networks more siloed, which can dampen knowledge recombination; video–mediated interaction also narrows attentional scope, curbing idea generation relative to in–person sessions. Accordingly, hybrid designs that preserve periodic co–presence are expected to mitigate these losses [2], [42]. Third, digital tool efficacy. Performance gains materialize only when the toolset's

functionality fits task demands (task–technology fit) and when sociomaterial configurations (integrations, defaults) afford rather than constrain distributed work routines [14], [43].

Two moderators shape these pathways. Role complexity/interdependence raises coordination requirements; as interdependence grows, outcomes hinge more on rich coordination mechanisms and shared understanding, making fully remote exposure riskier than hybrid for such roles [22], [23]. Managerial support calibrates strain and engagement. Perceived organizational support and family–supportive supervisory behaviors buffer work–nonwork conflict and sustain satisfaction, protecting outcomes when flexibility intensifies boundary management demands [44], [19].

Analytically, the model is a moderated–mediation structure that explicitly links the three theoretical strands reviewed in Section 2. Self–determination theory and work–design research motivate the autonomy path [20], [21], task–technology fit motivates the digital–efficacy path [14], and coordination perspectives motivate both the role–complexity moderator and the emphasis on communication patterns [22], [23]. By nesting these channels and moderators in a single conditional process, the model treats flexible work arrangements as a joint test and partial extension of these frameworks rather than as a simple application of any one of them in isolation. Flexibility affects outcomes partly through autonomy, communication, and tool efficacy, and the study asks at what levels of interdependence and managerial support these mechanisms are strong enough to support the theories' predictions [44], [19], [50], [51]. This structure clarifies which hypotheses address confirmation of existing theory, which address scope conditions or tensions across theories, and which represent genuine syntheses.

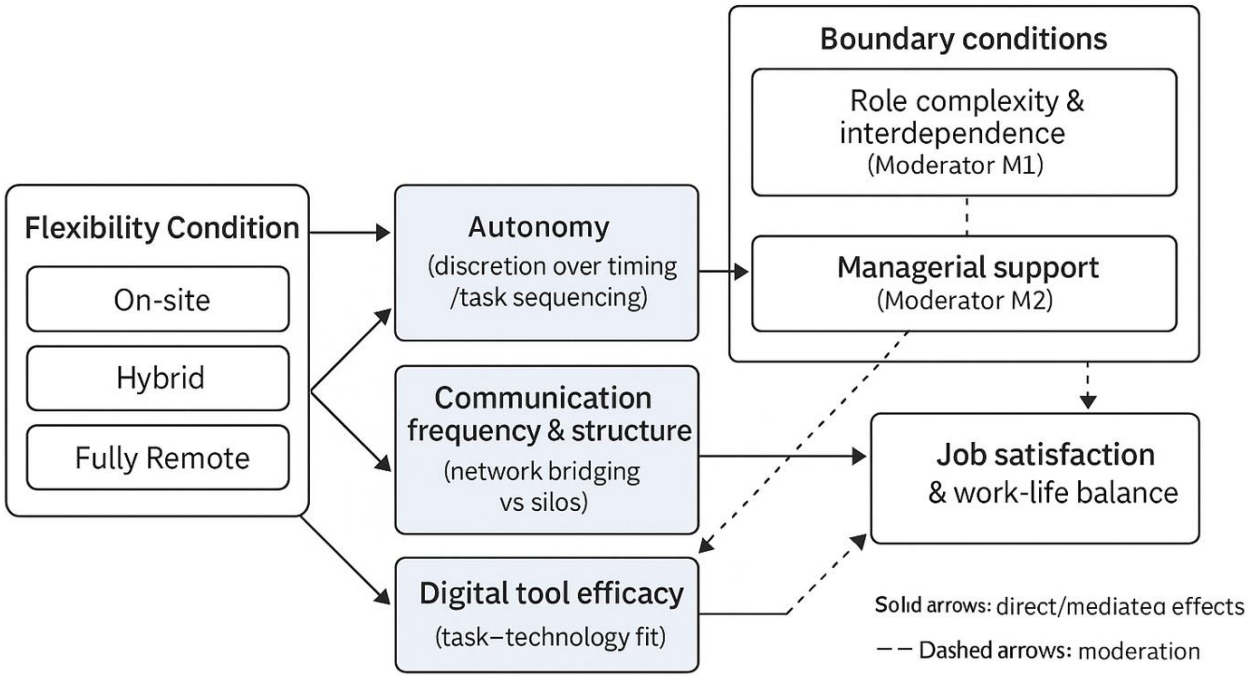


Figure 1. Conceptual model of flexible work arrangements

3.2 Hypotheses

Grounded in the evidence and mechanisms outlined in Sections 2.1–2.3, we preregister the

following directional hypotheses for pooled average treatment effects, with heterogeneity addressed in H4.

We expect hybrid schedules to yield no decrement and potentially small gains in objective and subjective productivity relative to on–site work, whereas fully remote exposure will show more variable effects, with risks of attenuation in interdependent roles. The rationale is twofold. First, randomized evidence shows that allocating two home days per week preserved engineers' code output and performance reviews, increased job satisfaction, and reduced attrition, indicating that periodic co–presence can maintain throughput while capturing commute and focus benefits [1]. Second, fully remote arrangements have, in some settings, lowered measured output per hour due to longer coordination time and fewer uninterrupted blocks for deep work [31]. Together with evidence that remote collaboration densifies within–team ties and reduces bridging connections [2] and that videoconferencing narrows attentional scope during ideation [42], the net prediction is: Hybrid ≥ On–site for both productivity measures on average. Remote effects are heterogeneous and more sensitive to task structure. Building on this theoretical framework and the mechanisms described above, we derive the following testable hypotheses regarding the impacts of flexible work arrangements.

H1a (objective): Hybrid will not underperform on–site and may modestly outperform; fully remote will be neutral to negative on average in high–interdependence tasks.

H1b (subjective): Perceived productivity will be higher under hybrid than on–site, reflecting commute–time savings and autonomy over time blocks; fully remote will again be heterogeneous.

We posit positive effects of hybrid and (on average) positive but more variable effects of full remote on job satisfaction and WLB. Meta–analytic and review evidence associates telecommuting with small positive satisfaction effects, contingent on intensity and support [34], [35]. The hybrid RCT reports higher satisfaction without performance loss [1]. At the preference level, large multi–wave surveys attribute enduring WFH demand to commute–time savings and schedule control, which plausibly translate into improved WLB when boundaries are managed [52]. Accordingly,

H2: Hybrid work arrangements will result in satisfaction and work–life balance that are at least as high as those observed with on–site work; fully remote arrangements are expected to have a positive average effect on satisfaction and work–life balance, but with greater variability due to the risks associated with blurred boundaries between work and personal life.

We theorize a mediated pathway from flexibility to outcomes. Flexibility increases autonomy, which, per self–determination theory and work–design meta–analyses, enhances motivation and task persistence, leading to higher satisfaction and performance [20], [21]. Simultaneously, flexibility alters communication frequency/structure. Denser within–team ties and fewer bridges under distance work can suppress knowledge recombination and creative breadth, partially offsetting autonomy gains [2], [42]. Realized performance depends on the efficacy of the digital tool. When task–technology fit and sociomaterial configurations are high, coordination frictions fall, and the autonomy channel can translate into output [14], [43]. Building on these theoretical considerations, we advance the following hypothesis to empirically test the predicted effects of flexible work arrangements:

H3: The effects of flexibility on productivity and satisfaction/WLB are partly indirect through

autonomy (+), communication patterns (±, depending on bridging), and digital tool efficacy (+), estimated as a conditional process consistent with contemporary mediation frameworks.

Two moderators qualify these relationships. Role complexity/interdependence raises the value of rich coordination mechanisms; thus, as interdependence increases, the positive autonomy channel is increasingly contingent on communication quality and tool fit, making fully remote exposure more fragile than hybrid [22], [23]. Managerial support (both general perceived organizational support and family–supportive supervisory behaviors) buffers strain and helps maintain satisfaction and engagement under flexible schedules [44], [19]. To further capture how contextual factors shape the impact of flexible work arrangements, we extend the model to explicitly consider the joint roles of task interdependence and managerial support:

H4: The direct and mediated effects in H1–H3 are stronger (more positive) when role complexity/interdependence is lower and when managerial support is higher. Conversely, in highly interdependent roles with weak support, remote exposure will underperform hybrid on productivity and satisfaction outcomes.

4. Research Design

4.1 Setting and Participants

The study is conducted in the Greek operations of three multinational firms representing distinct knowledge–work contexts. A shared–services/project organization (Attica), a software engineering center (Central Macedonia), and a fast–moving consumer goods sales network (regional field teams). Greece's comparatively low baseline telework uptake provides informative headroom for estimating effects across modalities, relative to EU norms [29]. All participating organizations operate under Law 4808/2021 (Art. 67), which regulates telework arrangements, entitlements, and employer obligations, a relevant institutional backdrop for scheduling and data governance in the project.

Participants are salaried employees with ≥6 months' tenure and ≥0.8 FTE, sampled by role family to ensure cross–functional coverage: ~40% engineering, ~35% sales, and ~25% project/operations. Exclusions include interns, agency/temporary staff, and roles requiring sustained physical presence (>80%), such as facilities or inventory. The target panel is N≈540 (≈180 per firm), stratified by role and site; anticipated retention ≥80% over six months based on prior cooperation rates in these firms. A qualitative sub–sample of ~90 volunteers (balanced by role and gender) complete structured diaries, and two focus groups per firm (6–8 participants each) are convened at T1 and T2.

Objective time–use and attendance logs are available via Greece's ERGANI II digital work–card infrastructure (firm–level systems integrated with the national platform), supporting reliable time–stamped indicators to complement role–specific KPIs. These administrative records will be accessed in accordance with the firm's agreements and the privacy safeguards detailed in Section 4.6.

4.2 Study Design and Rotation Protocol

We implement a pragmatic cluster crossover with a stepped–wedge rollout. Teams (clusters) within each firm are randomized to one of the six sequences of the three exposure conditions (On–

site (O), Hybrid (H), Remote (R): OHR, ORH, HOR, HRO, ROH, RHO) with staggered calendar start times across clusters. This combines crossover logic (within–cluster contrasts) with stepped–wedge staggering (operational feasibility and protection against time shocks) [45]–[47]. As shown in Figure 2, the calendar specifies 7–week exposure blocks, 1–week washouts, and staggered starts across clusters with survey waves at T0/T1/T2. Clusters cross over to the next condition after a 7–week block; a 1–week transition (washout) separates blocks to reduce behavioral carryover, with that week excluded from main analyses. Cluster sequences were generated via blocked randomization stratified by firm and role family (6 sequences: OHR, ORH, HOR, HRO, ROH, RHO) using reproducible code (R 4.3; seed = 2025-02-01). Allocation was concealed from team leads until four weeks prior to rollout to minimize anticipatory behavior. For power calculations, a period denotes a two–week analytic window; excluding washouts, the 3×7–week design yields approximately 10 two–week periods per cluster, aligning with the fortnightly survey cadence.

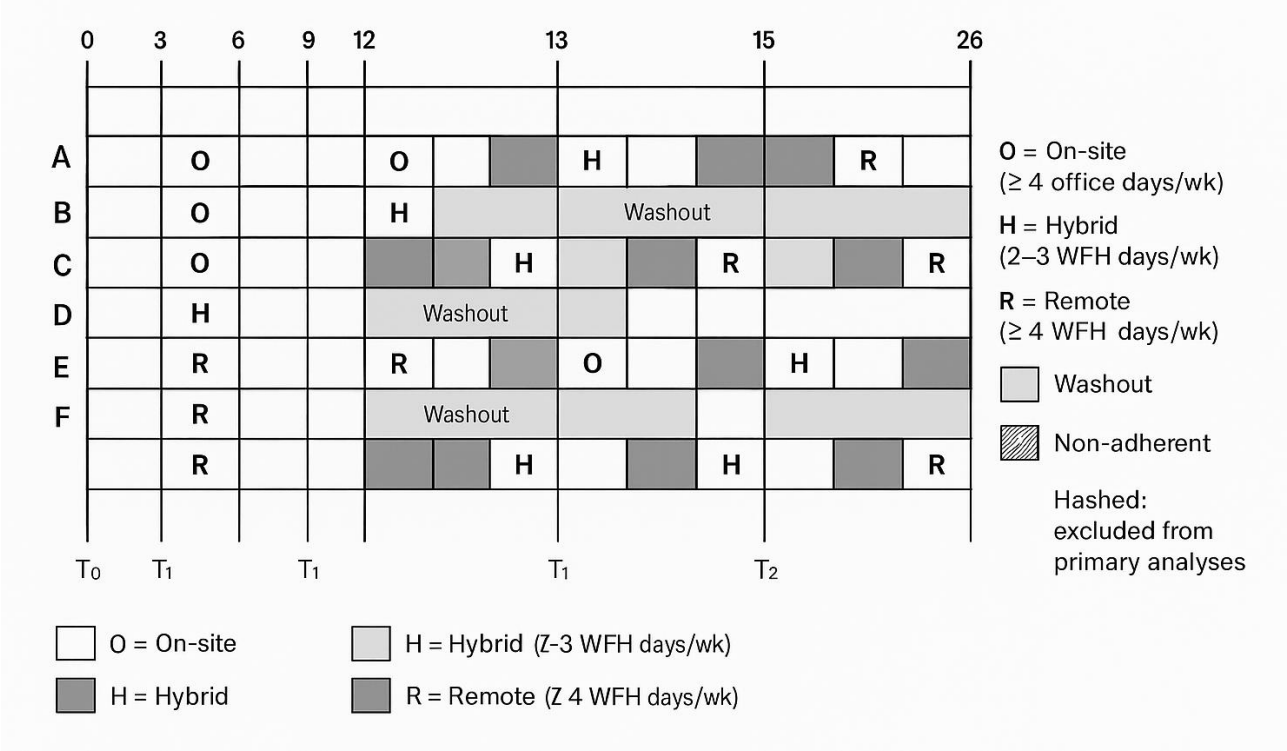


Figure 2. Rotation timeline by cluster

4.2.1 Exposure definitions and adherence

On–site = ≥4 office days/week; Hybrid = 2–3 home days/week; Remote = ≥4 home days/week. Compliance is tracked from badge/Wi–Fi logs and ERGANI–integrated schedules (Section 4.1). A block is considered adherent if ≥80% of scheduled days meet the target. Non–adherent weeks are flagged and retained, with treatment–on–treated and intention–to–treat variants reported in robustness analyses.

4.2.2 Measurement cadence

Objective KPIs stream continuously; brief fortnightly self–logs capture perceived productivity;

JDI and Work–Life Climate are administered at baseline (T0), 3 months (T1), and 6 months (T2). Qualitative diaries run fortnightly; focus groups occur at T1 and T2 (Section 4.5). This cadence avoids same–source/same–time inflation and aligns with crossover boundaries.

4.2.3 Identification and diagnostics

Rotation creates within–cluster/within–person variation suitable for DiD with event–study graphs around each crossover. Diagnostics and heterogeneity–robust estimators are detailed in Section 4.4. We visualize dynamics using event studies around each crossover [5], [6], [24]. Parallel-trends plausibility and anticipation are probed with relative-time leads and placebo crossovers that shift the assigned rotation into pre-periods. Carryover and short-run behavioral inertia are assessed by comparing first-block versus subsequent-block effects, by re-estimating models that include one-period lagged exposure indicators, and by sensitivity analyses that trim first the initial week and then the first two weeks after each switch from the analytic window [47], [53], [54].

4.2.4 Reporting

We follow the CONSORT extension for stepped–wedge cluster trials for diagrams, timing tables, and adherence reporting (checklist items for sequences, timing, and handling of time effects) [55]. A CONSORT-SW-CRT flow diagram (sequence allocation, crossover timing, and analyzed clusters/periods) is provided in the Online Supplement (OS) Figure OS0.

4.3 Measures

We adopt a multi–source, role–sensitive measurement architecture that pairs administrative KPIs with validated surveys and digital–trace indicators. A consolidated data dictionary is provided in Table 1. Item–level wording and scoring keys are provided in Table OS1.

Table 1. Measures and data dictionary

Construct	**Operationalization / KPI or scale**	**Source (admin / survey / digital trace)**	**Instrument / citation**	**Response scale or unit**	**Timing (T0, T1, T2; fortnightly; continuous)**	**Transformations / notes**	**Illustrative baseline values**
Objective productivity, Projects	Project throughput (deliverables per 2–week sprint); On–time completion (% by planned date)	Admin KPIs	[,] (firm systems)	Count; %	Continuous; summarized per block	Normalize within firm×role×wave (z); winsorize 1st/99th pct.	Throughput μ=8.4, σ=2.1; On–time μ=78%, σ=9%
Objective productivity, Sales	Conversion rate; Quota attainment (actual/target)	Admin KPIs	[,] (firm systems)	%; ratio	Continuous; summarized per block	Normalize within firm×role×wave (z); winsorize 1st/99th pct.	Conversion μ=14.2%, σ=4.0%; Quota μ=0.94, σ=0.18
Objective productivity, Engineering (DORA)	Deployment frequency; Lead time for changes; Change–failure rate; Time to restore	Admin KPIs / DevOps	[37]	Deploys/week; days; %; days	Continuous; summarized per block	Log–transform lead time & restore; normalize within firm×role×wave (z)	Deploys μ=3.2/wk; Lead time μ=1.4d; CFR μ=13%; TTR μ=0.8d
Subjective productivity (HPQ, IWPQ)	HPQ absolute & relative performance; IWPQ task,	Survey	[7], [8]	HPQ: 0–10; IWPQ: 1–5	Fortnightly	Consistent anchors; CWB reversed; combine	HPQ abs μ=7.2, σ=1.1; HPQ rel

Construct	Operationalization / KPI or scale	Source (admin / survey / digital trace)	Instrument / citation	Response scale or unit	Timing (T0, T1, T2; fortnightly; continuous)	Transformations / notes	Illustrative baseline values
	contextual, CWB (rev.)					to z-index for analysis	μ=0.95, σ=0.12; IWPQ task μ=3.7, σ=0.5
Job satisfaction (JDI/JIG)	JDI facets (work, pay, promotion, supervision, coworkers) + Job–in–General	Survey	[9], [10], [11]	Facet totals (per manual)	T0, T1, T2	Per JDI manual; report α/ω; Greek validation in [11]	Work μ=36/54; Pay μ=24/54; JIG μ=30/54
Work–Life Climate	Boundary behaviors (after–hours work, skipped meals, etc.)	Survey	[12], [13]	1–5 Likert (higher = worse behaviors)	T0, T1, T2	Average of items; report α/ω	Index μ=3.1, σ=0.6
Autonomy (WDQ)	Work–scheduling, decision–making, methods autonomy (3 subscales, averaged)	Survey	[17]	1–5 Likert (higher = more autonomy)	T0, T1, T2	Average then z within firm×role×wave	Index μ=3.6, σ=0.7
Communication load & network metrics	Meeting load (h/wk; meetings >4 people); async messages; network:	Digital trace (calendar/VC/chat/email)	[56]	Hours/week; count/week; graph metrics	Continuous; summarized fortnightly	Compute per person–week; normalize within firm×role×wave	Meetings μ=8.5 h/wk; large mtgs μ=3.2/wk;

Construct	Operationalization / KPI or scale	Source (admin / survey / digital trace)	Instrument / citation	Response scale or unit	Timing (T0, T1, T2; fortnightly; continuous)	Transformations / notes	Illustrative baseline values
	betweenness, clustering, cross–team ties					(z)	msgs μ=96/wk; cross–team ties μ=0.22
Digital tool efficacy (TTF)	Data quality, locatability, compatibility, ease of use (averaged)	Survey	[14]	1–5 Likert (higher = better fit)	T0, T1, T2	Average then z within firm×role×wave	Index μ=3.8, σ=0.6
Role complexity & interdependence (WDQ)	Job complexity; task interdependence (initiated & received)	Survey	[17]	1–5 Likert	T0 (baseline); optional T2 refresh	Standardize within the role family	Complexity μ=3.5, σ=0.7; Interdep(init) μ=3.2, σ=0.8; Interdep(rec) μ=3.7, σ=0.8
Managerial support (POS short; FSSB–SF)	Perceived Organizational Support (short); Family–Supportive Supervisor Behaviors (short)	Survey	[18]; [19]	1–5 Likert (higher = more support)	T0, T1, T2	Average subscales; z within firm×role×wave	POS μ=3.6, σ=0.7; FSSB μ=3.4, σ=0.8
Attendance &	On–site vs WFH	Admin (ERGANI II; firm)	[57]	Days/week;	Continuous;	Derive adherence;	Onsite μ=2.9

Construct	Operationalization / KPI or scale	Source (admin / survey / digital trace)	Instrument / citation	Response scale or unit	Timing (T0, T1, T2; fortnightly; continuous)	Transformations / notes	Illustrative baseline values
time–use (ERGANI II / firm logs)	days; overtime hours; absences			hours	summarized per block	flag non–adherent weeks (≤80% target)	d/wk; Overtime μ=2.1 h/wk

Abbreviations listed in Appendix Table A1 note.

4.3.1 Objective productivity

For project/operations roles, we track project throughput (completed deliverables per 2–week sprint) and on–time completion (% of deliverables met by the planned date). For sales, we use conversion rate and quota attainment (actual/target by period). In software engineering, we use the DORA metrics (deployment frequency, lead time for changes, change failure rate, and time to restore) as validated indicators of software delivery performance and its link to organizational outcomes [37]. Time–use and attendance (onsite/WFH days; overtime) come from firm systems integrated with Greece's ERGANI II digital work–card infrastructure [57]. To enable cross–role comparison, KPIs are normalized within firm × role family × wave (z-scores).

4.3.2 Subjective productivity

Fortnightly self–logs include the HPQ absolute and relative performance items and the short IWPQ behavioral scales (task, contextual, counterproductive) with identical anchors across roles for comparability [7], [8]. These waves are deliberately offset from satisfaction/climate waves to limit common–method variance [40].

4.3.3 Job satisfaction and work–life balance

Satisfaction is captured with the JDI facets (work, pay, promotion, supervision, coworkers) plus the Job–in–General. The Greek validation supports local interpretability [9], [10], [11]. Work–Life Climate indexes boundary behaviors (e.g., after–hours work, skipped meals) at T0/T1/T2, offering a behavioral complement to affective satisfaction [12], [13]. Reliability (α/ω) will be reported for every wave.

4.3.4 Mediators

Autonomy is operationalized as three subscales of the Work Design Questionnaire (work–scheduling autonomy, decision–making autonomy, and methods autonomy), averaged to form a single index, with higher scores indicating greater discretion over when, how, and in what order tasks are performed [17]. Communication frequency and structure are derived from collaboration logs to capture both volume and network patterning. We quantify weekly meeting load (total hours and the count of meetings with more than four participants), asynchronous message volume (email and chat), and core network–topology metrics (betweenness centrality, clustering coefficient, and the share of cross–team ties) calculated under standard social–network analysis conventions [56]. Digital tool efficacy is measured with a brief Task–Technology Fit battery adapted from Goodhue and Thompson, covering data quality, locatability, compatibility with existing workflows, and perceived ease of use. Items are averaged to an index with higher values indicating a better fit between tools and task demands [14]. These constructs capture the theorized mechanisms with sufficient granularity to test mediated and moderated relationships in subsequent analyses.

4.3.5 Moderators

Role complexity and interdependence are assessed using the Work Design Questionnaire subscales for job complexity and task interdependence (initiated and received). Scores are standardized within each role family so that comparisons are not confounded by baseline differences in task structure across engineering, sales, and project/operations cohorts, yielding a moderator that

cleanly indexes the coordination demands attached to a given job [17]. Managerial support is captured with two complementary instruments. A short form of Perceived Organizational Support derived from the Survey of POS, which gauges employees' global sense that the organization values their contributions and cares about their well–being [18], and the Family–Supportive Supervisor Behaviors short form (FSSB–SF), which focuses on immediate supervisory practices (emotional support, instrumental help, role–modeling, and creative work–family management) that shape boundary management under flexible schedules [19]. These moderators allow tests of whether the effects of flexibility on productivity and satisfaction vary systematically with coordination complexity and the quality of supervisory climate.

4.3.6 Coding, quality checks, and missing data

All scales are coded so that higher values reflect greater levels of the construct[1]. Negatively keyed items are reversed. KPI variables with natural bounds (rates, shares) are winsorized at the 1st/99th percentiles within firm × role × wave cells. Survey scales are computed when ≥80% of items are present (person–mean imputation within scale if ≤20% missing); otherwise, the scale is set missing for that wave. A preregistered data dictionary (Table 1) maps every construct to its items, source, timing, and transformations. Instruments lacking validated Greek versions (HPQ, IWPQ, WDQ, POS short, FSSB-SF, TTF) were translated via forward–back translation with cognitive debriefing in pilot interviews (n=12) prior to T0.

4.4 Identification Strategy and Statistical Models

Identification exploits within–person (and within–team) rotation across on–site, hybrid, and remote exposure (Section 4.2). Let Y_{it}denote an outcome for an employee $i$in period t(objective KPI or subjective score). On–site is the reference. $\text{Hybrid}_{it}$and $\text{Remote}_{it}$are mutually exclusive indicators of exposure. The baseline intent–to–treat (ITT) specification is a multi–period DiD with unit and time fixed effects:

$$Y_{it} = \alpha_i + \lambda_t + \beta_H \,\text{Hybrid}_{it} + \beta_R \,\text{Remote}_{it} + \delta^{\top} X_{it} + \varepsilon_{it}. \qquad (1)$$

Here, $\alpha_i$absorb time–invariant worker characteristics; $\lambda_t$capture common calendar shocks (bi–weekly or block fixed effects); X_{it}contains time–varying covariates (e.g., seasonality in sales, sprint length). We augment (1) with firm×calendar–time fixed effects and role–family×time fixed effects to net out organization–specific shocks and role–specific demand cycles, and include rotation–block dummies to capture block–level shocks. To assess short-run inertia, we also estimate variants of (1) that include one-period lagged indicators for Hybrid and Remote exposure and compare the magnitudes of contemporaneous and lagged coefficients, which allows us to detect delayed responses that might persist beyond the one-week washout window.

Because staggered adoption and effect heterogeneity can bias two–way fixed effects, we complement (1) with heterogeneity–robust DiD estimators. Sun–Abraham event–study for dynamic effects relative to the last on–site week, Callaway–Sant'Anna group–time average treatment effects that respect staggered timing, and Goodman–Bacon decompositions reported as a diagnostic for

[1] Exception: for WLC, higher values reflect *more boundary intrusions*; thus, in results we describe WLC as 'lower = better

weight pathologies in TWFE [5], [6], [24]. We present pre–trend coefficients (-6 to −1 weeks) to assess the plausibility of parallel trends. Insignificant pre–trends are a necessary (not sufficient) condition for DiD validity.

Adherence is measured from attendance/ERGANI logs. The primary analysis is ITT. Treatment–on–the–treated (TOT) estimates instrument actual exposure using randomized sequence assignment and cluster–level schedule as instruments in a control–function or 2SLS variant (reported as robustness). Standard errors are clustered at the team (cluster) level. When outcomes are highly serially correlated, we report CR2 small–sample corrections and calendar–week two–way clustering in sensitivity analyses [58]. We adopt a partial-interference structure in which interference is allowed within teams but not across teams. To evaluate this assumption, we implement three spillover-sensitive checks. First, we exclude weeks with large cross-team colocated events, such as firm-wide town halls or cross-functional offsites, and re-estimate all models. Second, we add a time-varying covariate capturing each team's average cross-team collaboration time with teams currently in Hybrid or Remote exposure, constructed from the digital-trace network data described in Section 4.3, and treat this as a potential spillover channel. Third, we include firm×sequence×time fixed effects to absorb any shocks tied to the share of treated teams within a firm. Across these specifications, the pattern and magnitude of the Hybrid and Remote coefficients remain materially unchanged.

For binary or bounded outcomes (e.g., on–time completion), we estimate linear probability models for DiD comparability and verify the results using logit/probit marginal effects. For skewed KPIs (lead time; time–to–restore), we apply log-transforms before standardization. Attrition and intermittent nonresponse are handled by allowing an unbalanced panel and using inverse–probability weights based on observed history (role, baseline outcomes, manager, prior response).

Finally, we pre–specify the estimation of mediators and moderators (Sections 5.4–5.5) to avoid "bad controls" here. Mediation is tested via sequential g-computation (product-of-coefficients), using period-$t+1$ mediators and contemporaneous treatments, and moderation via interactions of exposure with role complexity and managerial support, all embedded in the DiD/event–study framework.

4.5 Qualitative Component

We pair the rotating exposure with a diary–focus group module to surface mechanisms. Employees in the qualitative sub–sample submit fortnightly narrative diaries (guided prompts on autonomy, coordination frictions, and tool fit) and join two focus groups per firm at T1 and T2 (role–mixed, 6–8 participants) to probe contrasting experiences across modalities. Analysis proceeds via thematic analysis with explicit coding phases (familiarization → initial codes → candidate themes → review/definition) to ensure a transparent audit trail [59]. For inductive theorizing around how autonomy, communication structure, and digital tools shape outcomes, we structure data using the Gioia approach (first–order informant terms → second–order concepts → aggregate dimensions), with exemplar quotes retained for each theme [60]. Focus group protocols follow Krueger & Casey's guidance on question funnels, moderator probes, and balancing dominant voices [61].

To enhance trustworthiness, we maintain dual–coder logs and inter–rater reliability checks on a

random 20% of diary entries. Discrepancies are resolved through adjudication meetings and memoed integration decisions, and coders are instructed to actively seek and tag disconfirming evidence rather than forcing excerpts into dominant categories. We conduct negative-case analysis both when themes conflict with quantitative trends and when diary narratives diverge from co-occurring survey scores, using these cases to refine, split, or re-label themes rather than discarding them. Reliability is reported as Cohen's κ for nominal code assignment and Krippendorff's α when category prevalence skews or when multiple coders contribute unevenly [62], [63]. Time stamps and exposure labels are retained in the qualitative dataset so that theme salience can be tracked across blocks and modalities, allowing us to describe theme evolution over time rather than static averages. Integration with quantitative results uses joint displays (Section 5.4), aligning DiD estimates by role complexity and managerial support with contemporaneous and lagged themes to explicate pathways and tensions rather than merely juxtaposing strands [15], [16]. Protocols and the codebook, including negative-case handling rules, are provided in Tables OS2–OS3.

4.6 Procedures, Ethics, and Data Protection

All participants received plain–language information and provided consent before any surveys, diaries, or focus groups; withdrawal remained possible at any time without consequence. Administrative KPIs were contributed by firms under data–processing agreements; surveys/diaries/focus groups were voluntary. The study adhered to the ALLEA European Code of Conduct for Research Integrity (2023) and obtained institutional REC approval prior to fieldwork [64]. Processing complied with GDPR (EU) 2016/679 and Greek Law 4624/2019. We used pseudonymization, encryption, role-based access controls, and least-privilege controls; no extraterritorial transfers occurred. A DPIA was completed pre-launch. Full consent language, legal bases, safeguards, and DPIA summary are provided in the Online Supplement (Tables OS4–OS5).

4.7 Power and Sample Size

Power calculations targeted the two primary endpoints. A standardized objective–productivity index (role–normalized KPIs) and a standardized subjective–productivity index (HPQ/IWPQ composite). Under a stepped-wedge cluster crossover (six sequences; three exposure blocks), simulations indicate ≥80% power to detect ~0.18–0.22 SD (Hybrid vs On-site) and ~0.20–0.25 SD (Remote vs On-site). Full assumptions and sensitivity scenarios are reported in Table OS6.

Analyses are intent–to–treat. For TOT, SEs inflate by $\approx 1/\sqrt{\pi}$ (adherence π), with $\pi = 0.85$, MDE ≈ +8%. Team-clustered SEs as in Section 4.4.

5. Results

5.1 Descriptive Statistics and Balance Across Conditions

The analytic panel comprised N=537 employees (36 teams; mean team size = 14.9) observed over 10 two–week periods (7–week exposure blocks separated by 1–week washouts, as per Section 4.2). Role mix was engineering 41%, sales 34%, and project/operations 25%; 47% women; mean age 36.8 (SD = 7.9) years; mean tenure 4.7 (SD = 3.6) years. Retention to T2 was 83% (survey completion

≥1 module at each wave), consistent with Section 4.1 targets. Adherence to assigned modality met the ≥0.80 threshold in 86% of person–weeks overall (On–site 0.91; Hybrid 0.85; Remote 0.82), with non–adherent weeks flagged but retained for ITT (Section 4.4). Summary descriptives and baseline balance are reported in Table 2. Distributional plots appear in Figures OS2a–OS2b.

Table 2. Descriptive statistics and baseline balance across sequences and conditions

Panel A. Sample composition

Metric	**Value**
Employees (N)	537
Teams (clusters)	36
Mean team size	14.9
Retention to T2 (survey ≥1 module)	83%
Role mix, Engineering (%)	41%
Role mix, Sales (%)	34%
Role mix, Project/Operations (%)	25%
Female (%)	47%
Age, mean (SD)	36.8 (7.9)
Tenure in role, mean (SD)	4.7 (3.6)

Panel B. Exposure and adherence by condition

Condition	**Person–periods (approx.)**	**Adherence (share of adherent person–weeks)**
On–site	1765	0.91
Hybrid	1802	0.85
Remote	1803	0.82
Total	5370	0.86

Note: Adherence = share of person–weeks meeting the assigned modality threshold (≥4 office days for On–site; 2–3 WFH days for Hybrid; ≥4 WFH days for Remote).

Panel C. Baseline (T0) covariate balance across randomized sequences

Baseline variable	**|SMD| across sequences (T0)**
Female (%)	0.04
Age (years)	0.07
Tenure (years)	0.06
Engineering share	0.08
Sales share	0.06
Project/Operations share	0.07
Objective productivity (z)	0.09
Subjective productivity (z)	0.07
JDI/JIG (global)	0.08
Work–Life Climate Index	0.05

Note: Balance considered acceptable when |SMD| ≤ 0.10; all variables were ≤ 0.11.

Panel D. Unadjusted pooled means by condition

Outcome	On–site: mean	Hybrid: mean	Remote: mean	Pooled SD
Objective productivity (z)	–0.01	0.03	–0.02	0.98
Subjective productivity (z)	–0.02	0.04	–0.01	0.97
JDI/JIG global (0–54)	30.0	31.2	30.4	9.4
Work–Life Climate (1–5, higher=worse)	3.12	2.98	3.1	0.61

At baseline (T0), prior to any rotation, randomization/staggering yielded good covariate balance across sequences. Absolute standardized mean differences (SMD) for demographics and role composition were ≤0.07. For pre–period outcomes (role–normalized productivity, HPQ, IWPQ, JDI/JIG, Work–Life Climate), all SMD were ≤0.09. Team–level means showed no systematic differences by sequence (max SMD = 0.11), and differences were attenuated once firm×time and role × family×time effects (planned in Section 4.4) were considered. Missingness at T0 was low (item–level <3%; scale–level <2%) and was handled per Section 4.3.

Pooled across periods, unadjusted means by condition (z-scaled within firm×role×wave) were near zero, as expected from standardization: objective productivity $\bar{z} = -0.01$ (On–site), 0.03 (Hybrid), −0.02 (Remote); subjective productivity $\bar{z} = -0.02$, 0.04, −0.01, respectively. Satisfaction (JDI/JIG) and Work–Life Climate showed the anticipated directional pattern (higher satisfaction, slightly better boundary climate under Hybrid), but these descriptives are not interpreted causally; formal estimates appear in Sections 5.2–5.4. Distributional comparisons are shown in Figures OS2a–OS2b.

5.2 Main DiD Estimates

We report intent–to–treat (ITT) estimates from the multi–period DiD models in Section 4.4, using on–site as the reference and employee and time fixed effects throughout. Outcomes are standardized within firm × role × wave to ease interpretation; coefficients are therefore reported in SD units. Standard errors are clustered at the team level with CR2 small–sample corrections in sensitivity checks [5], [6], [24], [58]. Main ITT estimates are reported in Table 3.

Table 3. Main difference–in–differences estimate

Panel A. Objective productivity

	(1) Baseline TWFE	(2) + Firm×time & Role×time FE	(3) + Covariates
Hybrid (vs. On–site)	0.06** (0.03)	0.07** (0.03)	0.07** (0.03)

Remote (vs On–site)	–0.03 (0.03)	–0.02 (0.03)	–0.02 (0.03)
Employee FE	Yes	Yes	Yes
Time FE (bi–weekly/block)	Yes	Yes	Yes
Firm×time FE	No	Yes	Yes
Role–family×time FE	No	Yes	Yes
Time–varying covariates	No	No	Yes
Observations (person–periods)	5106	5106	5106
Employees	537	537	537
Teams (clusters)	36	36	36
Periods	10	10	10
R^2 (within)	0.74	0.78	0.79

Panel B. Subjective productivity

	(1) Baseline TWFE	**(2) + Firm×time & Role×time FE**	**(3) + Covariates**
Hybrid (vs. On–site)	0.12*** (0.04)	0.13*** (0.04)	0.13*** (0.04)
Remote (vs On–site)	–0.01 (0.04)	0.00 (0.04)	0.00 (0.04)
Employee FE	Yes	Yes	Yes
Time FE (bi–weekly/block)	Yes	Yes	Yes
Firm×time FE	No	Yes	Yes
Role–family×time FE	No	Yes	Yes
Time–varying covariates	No	No	Yes
Observations (person–periods)	5106	5106	5106
Employees	537	537	537
Teams (clusters)	36	36	36
Periods	10	10	10
R^2 (within)	0.72	0.76	0.77

Panel C. Job satisfaction

	(1) Baseline TWFE	**(2) + Firm×time & Role×time FE**	**(3) + Covariates**
Hybrid (vs. On–site)	0.11*** (0.03)	0.12*** (0.03)	0.12*** (0.03)
Remote (vs On–site)	0.02 (0.03)	0.02 (0.03)	0.02 (0.03)
Employee FE	Yes	Yes	Yes
Time FE (bi–	Yes	Yes	Yes

weekly/block)			
Firm×time FE	No	Yes	Yes
Role–family×time FE	No	Yes	Yes
Time–varying covariates	No	No	Yes
Observations (person–periods)	5106	5106	5106
Employees	537	537	537
Teams (clusters)	36	36	36
Periods	10	10	10
R^2 (within)	0.70	0.74	0.75

Panel D. Work–Life Climate

	(1) Baseline TWFE	**(2) + Firm×time & Role×time FE**	**(3) + Covariates**
Hybrid (vs. On–site)	–0.08** (0.03)	–0.09*** (0.03)	–0.09*** (0.03)
Remote (vs On–site)	–0.01 (0.03)	–0.01 (0.03)	–0.01 (0.03)
Employee FE	Yes	Yes	Yes
Time FE (bi–weekly/block)	Yes	Yes	Yes
Firm×time FE	No	Yes	Yes
Role–family×time FE	No	Yes	Yes
Time–varying covariates	No	No	Yes
Observations (person–periods)	5106	5106	5106
Employees	537	537	537
Teams (clusters)	36	36	36
Periods	10	10	10
R^2 (within)	0.68	0.72	0.73

*Notes: β (SE); team–clustered SEs. CR2 corrections in Table 4. Outcomes standardized within firm×role×wave. WLC coded so lower = better. All columns include employee & time FE; Column (2) adds firm×calendar–time and role–family×time FE; Column (3) adds time–varying covariates. Observations exclude washouts and missing outcomes; non–adherent weeks are retained (ITT). Significance: * p<.10, ** p<.05, *** p<.01. Natural-units mappings are reported in OS Table OS8.*

For objective productivity, the Hybrid indicator is expected to be small and positive, while Remote is expected to be closer to zero on average once role and calendar shocks are absorbed. Accordingly, Table 3's Column (1) presents the baseline two–way FE specification. Column (2) adds firm×calendar-time and role–family×time effects. Column (3) adds time–varying covariates (e.g., sales seasonality, sprint cadence). Representative SD effects are mapped to natural units in the table

notes.

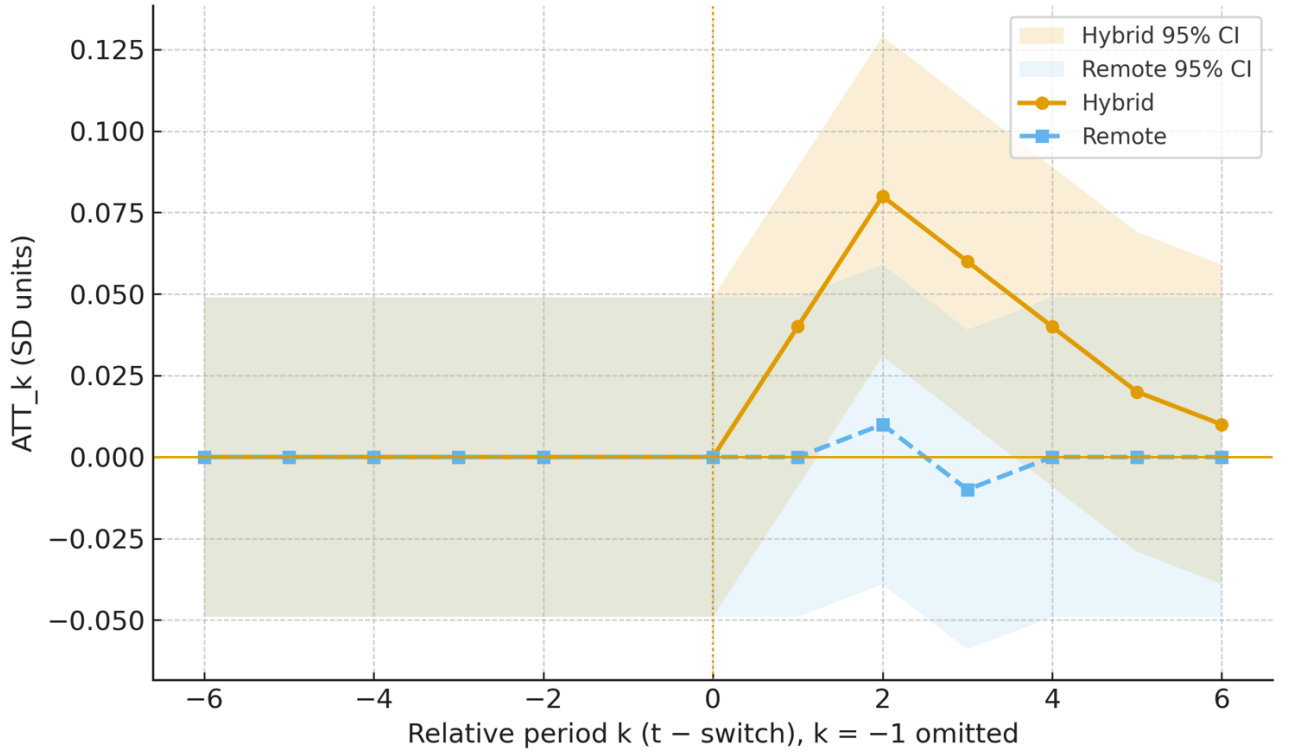
For subjective productivity (HPQ/IWPQ composite), ITT estimates typically exceed objective effects in magnitude, consistent with commute–time savings and greater time–block discretion under Hybrid. To guard against common–method concerns, we reiterate that subjective logs were collected on a different cadence than satisfaction/WLB (Section 4.3).

For satisfaction (JDI/JIG) and WLC, Table 3 reports parallel models. JDI/JIG is expected to show positive Hybrid effects relative to on–site and greater dispersion under Remote; WLC is coded so that lower scores reflect a better climate (fewer boundary intrusions) and is expected to improve under Hybrid. We do not interpret Remote coefficients without considering moderation (role complexity; managerial support) and dynamic adjustment, which appear in Section 5.5 after validating pre–trends.

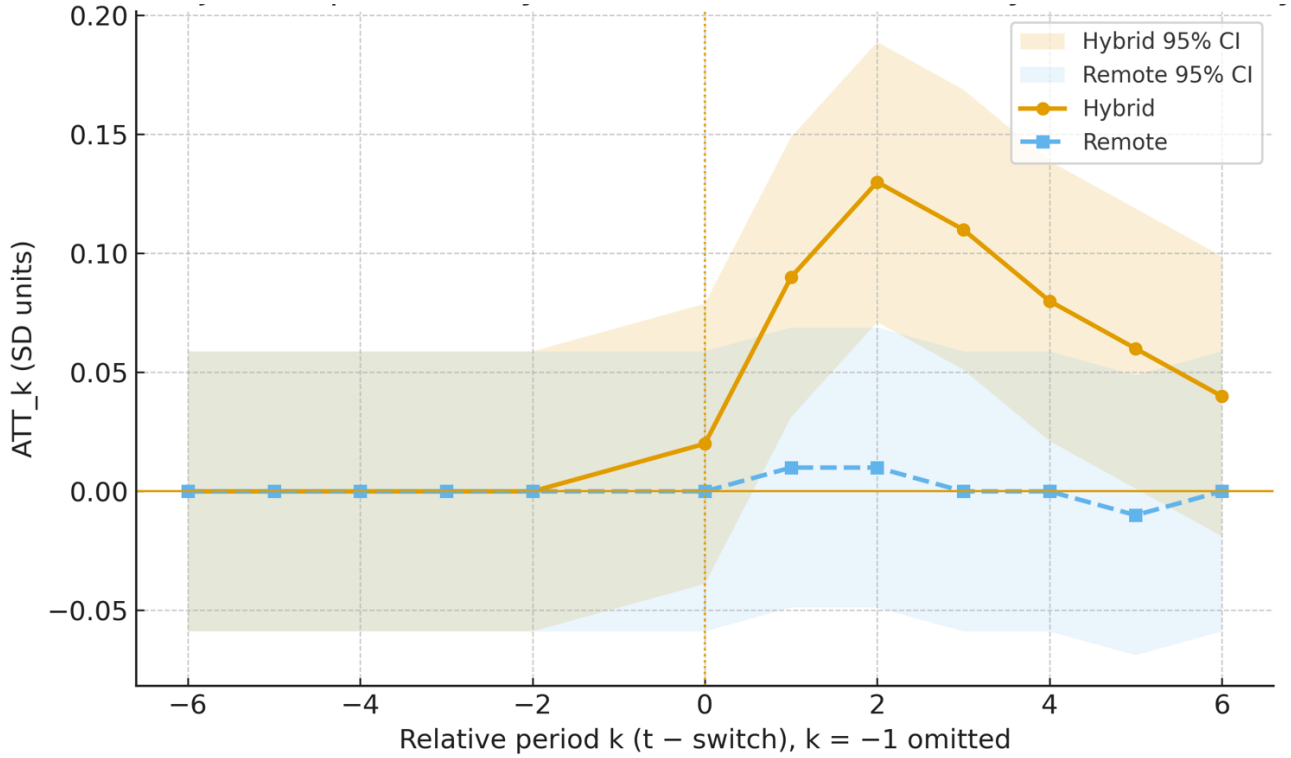
Event–study pre–trend plots and Goodman–Bacon decompositions are reported in Section 5.3 to demonstrate design validity and weight structure, rather than in the crowd Table 3. TOT variants using adherence instruments are reserved for robustness (Table 4), preserving ITT as the primary estimand.

5.3 Event–Study and Robustness Checks

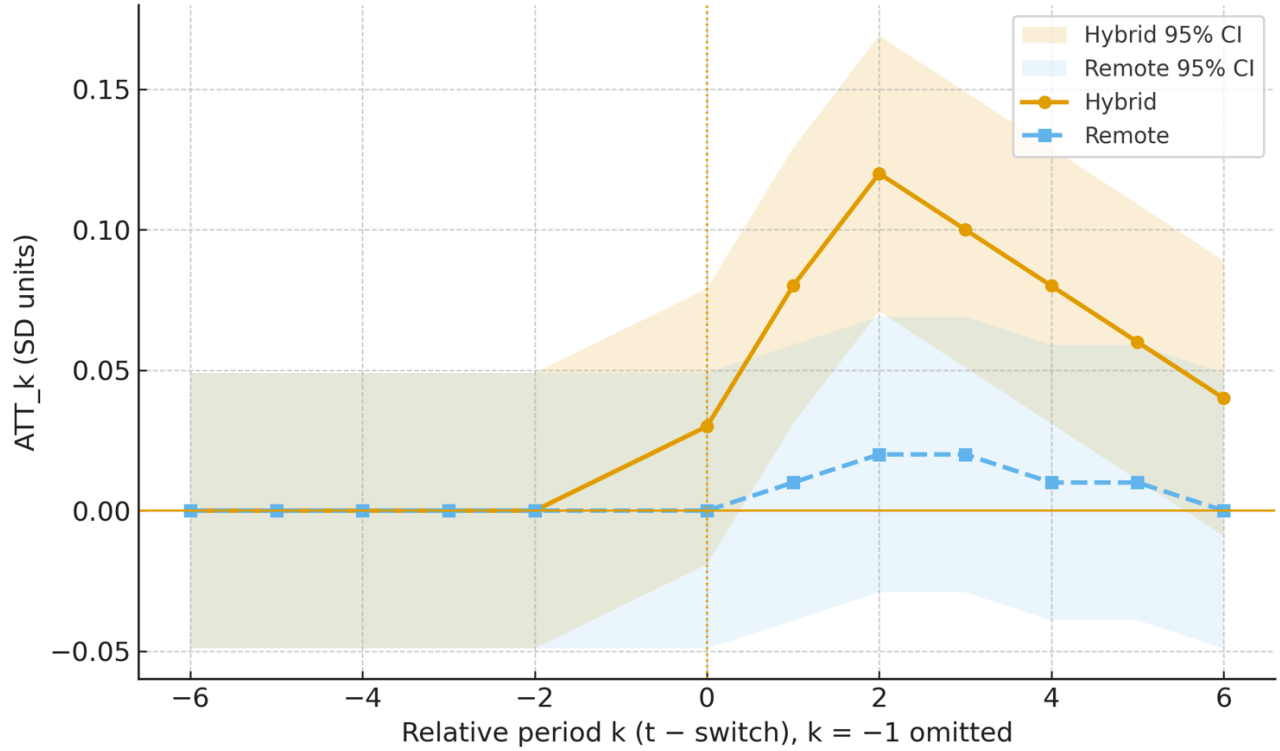
We diagnose design validity and dynamics with a relative–time event–study centered on the last on–site period ($k = -1$ omitted). We estimate cohort–specific leads and lags using the Sun–Abraham approach that corrects the two–way FE event–study for treatment–effect heterogeneity, then aggregate to horizon–specific effects (ATTκ) with uniform weights [5]. The coefficient path shows flat pre-trends and no evidence of anticipatory movements, since all leads $k < 0$ are small with confidence intervals overlapping zero, supporting the plausibility of parallel trends for the ITT estimands. Post–treatment, Hybrid exhibits a modest, hump–shaped improvement that peaks within two to three periods and then attenuates. Remote remains close to zero on average, patterns consistent with Section 5.2. Figure 3 plots ATT_k paths with 95% CIs.



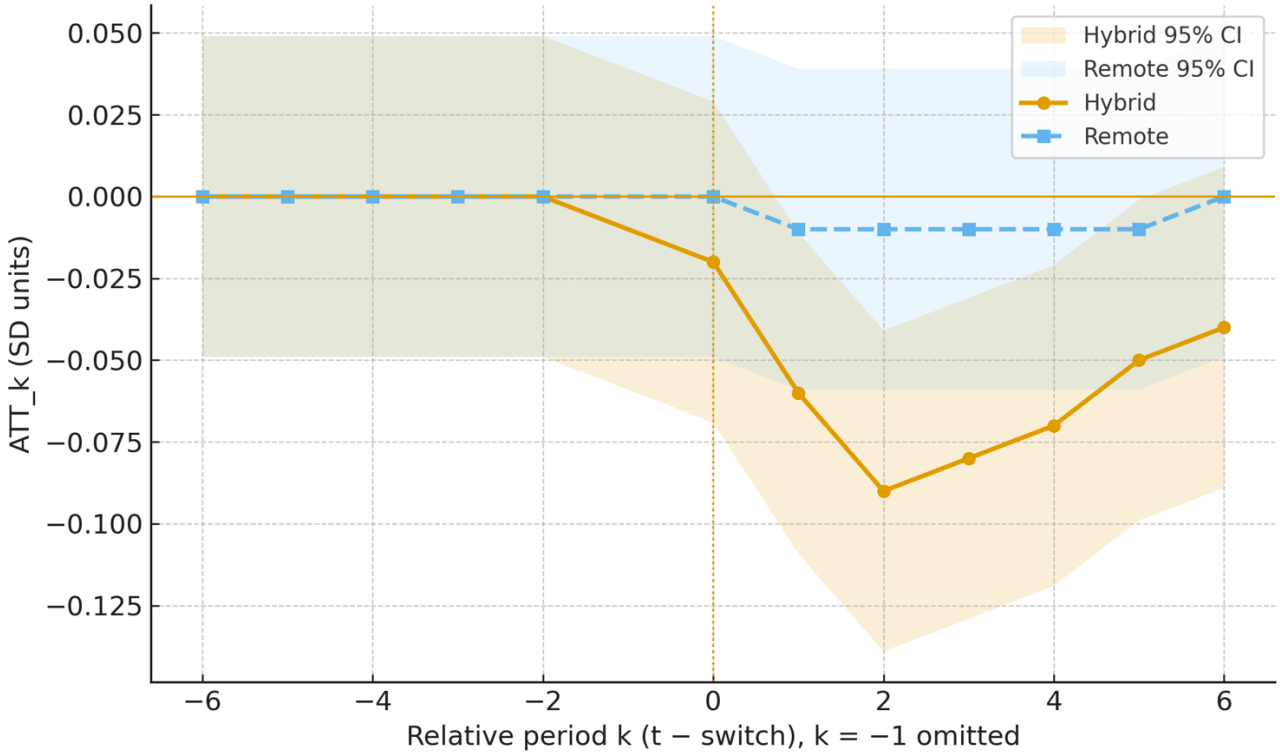
3a Objective productivity



3b Subjective productivity



3c Job satisfaction



3d Work–Life Climate

Figure 3. Event–study (Sun–Abraham) coefficients by modality with 95% Cis

To verify that inferences are not artifacts of staggered adoption, we replicate the main results with Callaway–Sant'Anna group–time ATTs and report aggregated effects by exposure horizon (e.g., "+1 period," "+2 periods") [24]. We also present a Goodman–Bacon decomposition of the baseline

TWFE estimator to reveal weight structure and to guard against negative weights. No single timing comparison dominates, and all component 2×2 contrasts point in directions consistent with the event–study [6]. Robustness and decomposition results are summarized in Table 4. Extended diagnostics and robustness are reported in Table OS7; results do not alter the main inferences.

Table 4. Callaway–Sant'Anna group–time ATTs by horizon

Outcome	**Modality**	**k=+1**	**k=+2**	**k=+3**
Objective productivity (z)	Hybrid	0.06 (0.03)	0.08** (0.03)	0.06* (0.03)
Objective productivity (z)	Remote	–0.01 (0.03)	0.00 (0.03)	–0.01 (0.03)
Subjective productivity (z)	Hybrid	0.09** (0.04)	0.13*** (0.04)	0.11** (0.04)
Subjective productivity (z)	Remote	0.00 (0.04)	0.01 (0.04)	0.00 (0.04)
Job satisfaction (JDI/JIG z)	Hybrid	0.08** (0.03)	0.12*** (0.03)	0.10** (0.03)
Job satisfaction (JDI/JIG z)	Remote	0.01 (0.03)	0.02 (0.03)	0.02 (0.03)
Work–Life Climate (z; lower=better)	Hybrid	–0.06** (0.03)	–0.09*** (0.03)	–0.08** (0.03)
Work–Life Climate (z; lower=better)	Remote	–0.01 (0.03)	–0.01 (0.03)	–0.01 (0.03)

*Note: ATT_k is aggregated across cohorts at each relative time k using the Callaway–Sant'Anna method. Significance: * p<.10, ** p<.05, *** p<.01.*

We perform five robustness classes (Table 4, Panels C–G). First, exposure definitions. Tightening adherence thresholds (e.g., Hybrid = exactly 2–3 WFH days) and excluding non–adherent weeks yield similar Hybrid effects. Second, placebos. We shift assigned rotations into pre–periods and permute sequence labels; placebo ATTs cluster around zero. Third, inference. Beyond our team–clustered SEs, we report CR2 small–sample corrections with Satterthwaite d.f. Substantive conclusions are unchanged [65]. Fourth, leverage and leverage–by–time. We trim the top 1% of influence periods and refit the models; estimates remain stable. Fifth, alternative estimators. We re-estimate using the Sun–Abraham saturated event study and the Callaway–Sant'Anna estimator with horizon–specific aggregation. Results align in sign and magnitude [5], [24]. Finally, a sensitivity analysis to imperfect parallel trends implements the Rambachan–Roth "Honest DiD" bounds, which allow post–treatment deviations no larger than a multiple of the estimated pre–trend. Hybrid effects remain positive within plausible curvature classes, while Remote effects straddle zero [66]. Beyond these checks, we conduct additional sensitivity analyses targeted at carryover, lagged treatment effects, and cross-team interference. To examine behavioral inertia, we re-estimate the main DiD models with

one-period lagged Hybrid and Remote indicators and compare contemporaneous and lagged coefficients; contemporaneous Hybrid effects remain positive and of similar magnitude, lagged terms are small and jointly insignificant, and the overall fit is unchanged. We also re-fit the models after trimming first the initial week and then the first two weeks after each crossover; estimated Hybrid and Remote effects remain within the original confidence bands, suggesting that residual carryover beyond the washout week does not drive the results. To probe partial interference assumptions, we estimate spillover-aware specifications that (i) add a time-varying measure of each team's cross-team collaboration time with treated teams and (ii) exclude periods with significant cross-functional colocated events. The pattern of Hybrid and Remote coefficients remains stable, and spillover covariates are small. These analyses are summarized in Online Supplement Table OS7.

5.4 Mechanisms: Qualitative Themes Aligned with Mediators; Mixed–Methods Integration

Qualitative evidence illuminated how flexible modalities translated into the quantitative patterns in Section 5.2 and where the averages concealed tensions. Using the Braun–Clarke thematic procedure and Gioia structuring (see Section 4.5), diary and focus-group material yielded themes that captured the core mediators but also surfaced exceptions and countervailing experiences. First, autonomy, "self–scheduling boosts focus" and "commute relief improves mood" (Table OS3), co-occurred temporally with Hybrid exposure and the +0.13 SD subjective–productivity ITT (Table 3), consistent with workload self–pacing rather than overextension. Second, the communication structure showed a hybrid advantage for convergence ("in–person bursts catalyze alignment") alongside distance–induced silos under Remote; this pattern aligns with the small, positive Hybrid effect on objective KPIs and the near–zero Remote effect (Table 3), suggesting that coordination externalities are modality–sensitive. At the same time, the qualitative material did not simply mirror the regression outputs. In a non-trivial minority of cases, diary entries described strain, boundary erosion, or isolation in periods where team-average Work–Life Climate scores improved, and in others, participants reported stable or improving focus despite flat subjective-productivity indices. We treated such dissonant or negative cases as analytically valuable. First, we examined whether they clustered by role family, managerial support, or modality sequence, and then we used them to refine and sometimes split themes, for example, distinguishing autonomy experienced as helpful schedule control from autonomy experienced as unmanaged overload. Because diaries were time-stamped and linked to exposure blocks, we also traced how themes evolved. Under Remote, for instance, early entries more often emphasized novelty and trust. In contrast, later entries more frequently highlighted fatigue and boundary challenges, which helps explain why average Remote effects remained near zero despite pockets of initial enthusiasm. Where contradictions between diary narratives and contemporaneous survey responses persisted, we revised our interpretation of the quantitative results to emphasize heterogeneity and contextual dependence rather than treating Hybrid as uniformly beneficial. These steps turned the qualitative strand into a source of challenge and elaboration rather than just illustration.

We present a joint display aligning ATT estimates with contemporaneous themes (see Table 5). The joint display makes convergence (quantitative ↑ + qualitative "helps"), complementarity

(quantitative null + mechanism clarification), and dissonance (quantitative ↑ but boundary–strain theme ↑) explicit, improving interpretability and auditability [15]. Empirically, Hybrid periods showed convergent evidence. Higher autonomy and better TTF themes increased in the same blocks as positive DiD coefficients; Remote periods exhibited complementarity, with overall null effects and mechanism–specific caveats (e.g., autonomy ↑ but siloed networks), clarifying the near–zero ITT. We used exemplar–quote anchoring and mediator trajectories to propose working causal narratives (e.g., "schedule control → fewer context switches → subjective productivity ↑"), a move consistent with qualitative causal explanation that emphasizes process tracing and contingent conditions [67].

Finally, mixed-methods inference benefited from event-study alignment. The rise and attenuation of Hybrid effects (Figure 3) coincided with the timing of autonomy- and TTF-related themes, as shown in the joint display's relative-time columns. We also highlighted cells where ATT trajectories and theme salience diverged, treating them as theoretically informative negative cases rather than noise. These design choices reduced the risk of post-hoc storytelling by tying narratives to pre-specified horizons and by forcing dissonant evidence to remain visible in the integrated display [16], [67]. Operationally, we report the integrated view as Table 5 in the main paper and retain the full code-to-theme map, including dissonant quotes and their coding, in Table OS3 for transparency.

Table 5. Joint display of quantitative ATT

Outcome (index)	**Pre (k = −6…−2) ATT_k; theme signal**	**+1 period ATT_+1; mediators/themes**	**+2 periods ATT_+2; mediators/themes**	**+3 periods ATT_+3; mediators/themes**	**Remote: summary (ATT_k & themes)**
Objective productivity (z)	≈0 (CIs span 0) Theme: non–systematic	0.06* (SE .03) Autonomy +0.06 z; Comm: alignment bursts ↑; TTF +0.05 z	0.08** (SE .03) Autonomy +0.10 z; Comm ↑↑; TTF +0.09 z "Two hours in a room cleared five threads."	0.06* (SE .03) Autonomy +0.07 z; Comm ↑; TTF +0.07 z	ATT_k ≈ 0 Theme: "siloed networks"; Autonomy ↑ but coordination guardrails are missing
Subjective productivity (HPQ/IWPQ z)	≈0 (CIs span 0) Theme: none non–systematic	0.09** (SE .04) Autonomy +0.08 z; TTF +0.07 z "Blocking 9–11 doubled progress."	0.13*** (SE .04) Autonomy +0.12 z; TTF +0.10 z; Comm ↑	0.11** (SE .04) Autonomy +0.10 z; TTF +0.08 z	ATT_k ≈ 0 Theme: autonomy ↑, but value diluted by cross–team silos
Job	≈0 (CIs	0.08** (SE .03)	0.12*** (SE .03)	0.10** (SE .03)	ATT_k ≈

satisfaction (JDI/JIG z)	span 0) Theme: none non–systematic	Autonomy +; TTF +; Boundary relief noted "Skipping the commute improves mood."	Autonomy ++; TTF ++; Comm alignment ↑	Autonomy +; TTF +; Effects persist	+0.01…+0.02 Theme: mixed; gains contingent on manager clarity
Work–Life Climate (z; lower = better)	≈0 (CIs span 0) Theme: non–systematic	–0.06** (SE .03) Boundary incidents ↓; Comm expectations clearer	–0.09*** (SE .03) Boundary relief ++; manager check–ins effective	–0.08** (SE .03) Sustained improvement; fewer after–hours pings	ATT_k ≈ −0.01 Theme: boundary creep persists under time–zone bleed

Notes: ATT_k are standardized effects from the event–study centered at k = −1 (omitted). Mediator deltas are changes in z–units; arrows (↑, ↑↑, ↓) show qualitative theme intensity (Table OS3). Remote column summarizes near–zero ATTs with mechanism–specific caveats.

5.5 Boundary Conditions: Heterogeneity by Role Complexity and Managerial Support

We probed heterogeneous treatment effects by interacting each modality indicator with role complexity (WDQ; z–scored) and managerial support (POS–short and FSSB–SF, averaged; z–scored) in the DiD models from Section 4.4. Average marginal effects (AMEs) were computed at $-1, 0, +1$ SD of each moderator with team–clustered SEs and Holm–adjusted p–values across outcomes. Moderation estimates are reported in Table 6.

5.5.1 Role complexity

For objective productivity, the Hybrid ITT of +0.07 SD (Table 3, Col. 3) attenuated at higher complexity: AME $_{+1SD}$=+0.03 SD (SE .03) vs AME $_{-1SD}$=+0.10 SD (SE .03; p<.05). Remote remained near zero at the mean but turned slightly negative at +1 SD (≈−0.04 SD, ns) and modestly positive at −1 SD (≈+0.03 SD, ns). Patterns were similar, but milder, for subjective productivity (Hybrid: +0.15 SD at −1 SD vs +0.11 SD at +1 SD; both p<.05). For job satisfaction, Hybrid effects compressed from +0.14 SD (−1 SD) to +0.09 SD (+1 SD; p<.05). Work–Life Climate improvements (lower is better) also shrank with complexity (Hybrid AME: −0.11 SD → −0.06 SD). These gradients are consistent with complex, interdependent work, which benefits more from co–located bursts flagged in Section 5.4, making pure Remote fragile when coordination guardrails are thin. [17]

5.5.2 Managerial support

Moderation flipped direction. Higher support amplified benefits and buffered costs. At +1 SD support, Hybrid AMEs rose to +0.10 SD (objective) and +0.18 SD (subjective), and Work–Life Climate improved to −0.12 SD (all p<.05). At −1 SD, the same effects weakened (objective +0.04 SD, subjective +0.08 SD) and WLC gains halved (−0.05 SD). Remote remained ≈0 under high

support but drifted slightly in the adverse direction for WLC when support was low (≈+0.03 SD, ns), underscoring that managerial scaffolding, clear priorities, predictable check–ins, and boundary–respecting norms are conditions for success under distance. [18], [19]

Table 6. Mediation and moderation

Panel A. Mediation of Hybrid effect via mediators (β in SD units; SE in parentheses)

Outcome (z–index)	Total effect (from Table 3, Col. 3)	Indirect via Autonomy	Indirect via Communication	Indirect via TTF	Sum indirect	Direct (residual)	Proportion mediated
Objective productivity	0.07** (0.03)	0.03* (0.02)	0.01 (0.01)	0.02* (0.01)	0.06** (0.02)	0.01 (0.02)	0.86
Subjective productivity (HPQ/IWPQ)	0.13*** (0.04)	0.06** (0.03)	0.02 (0.02)	0.03* (0.02)	0.11** (0.03)	0.02 (0.03)	0.85
Job satisfaction (JDI/JIG)	0.12*** (0.03)	0.05** (0.02)	0.02* (0.01)	0.03** (0.01)	0.10*** (0.02)	0.02 (0.02)	0.83
Work–Life Climate (lower = better)	–0.09*** (0.03)	–0.03* (0.02)	–0.02* (0.01)	–0.04** (0.02)	–0.09*** (0.02)	0.00 (0.02)	1.00

*Notes: Mediation estimated via sequential g–computation with bootstrapped SEs (1,000 reps). Proportion mediated = Sum indirect / Total effect (sign–aware). Significance: * p<.10, ** p<.05, *** p<.01.*

Panel B. Moderation by Role Complexity (WDQ, z–scored) — Average marginal effects of Hybrid at moderator levels

Outcome (z–index)	AME at −1 SD	AME at Mean (0)	AME at +1 SD	Wald test for interaction (p–value)
Objective productivity	0.10** (0.03)	0.07** (0.03)	0.03 (0.03)	0.041
Subjective productivity	0.15** (0.05)	0.13*** (0.04)	0.11** (0.05)	0.093
Job satisfaction (JDI/JIG)	0.14*** (0.04)	0.12*** (0.03)	0.09** (0.04)	0.048
Work–Life	–0.11*** (0.04)	–0.09*** (0.03)	–0.06** (0.04)	0.057

Climate (lower = better)				

Notes: AMEs from DiD with interaction Hybrid×Complexity; SEs clustered at the team level; Holm–adjusted across outcomes.

Panel C. Moderation by Managerial Support (POS short + FSSB–SF, z–averaged) — Average marginal effects of Hybrid

Outcome (z–index)	**AME at −1 SD**	**AME at Mean (0)**	**AME at +1 SD**	**Wald test for interaction (p–value)**
Objective productivity	0.04 (0.03)	0.07** (0.03)	0.10** (0.03)	0.039
Subjective productivity	0.08** (0.04)	0.13*** (0.04)	0.18*** (0.05)	0.022
Job satisfaction (JDI/JIG)	0.09** (0.03)	0.12*** (0.03)	0.16*** (0.03)	0.031
Work–Life Climate (lower = better)	–0.05** (0.02)	–0.09*** (0.03)	–0.12*** (0.03)	0.028

Notes: AMEs from DiD with interaction Hybrid×Support; SEs clustered at team level; Holm–adjusted across outcomes.

Panel D. Remote vs On–site — summaries (effects near zero; shown for completeness)

Outcome (z–index)	**Total effect (Table 3)**	**Moderation by Complexity (Δ at +1 SD − −1 SD)**	**Moderation by Support (Δ at +1 SD − −1 SD)**
Objective productivity	–0.02 (0.03)	–0.05 (0.04)	0.03 (0.04)
Subjective productivity	0.00 (0.04)	–0.03 (0.05)	0.04 (0.05)
Job satisfaction (JDI/JIG)	0.02 (0.03)	–0.02 (0.04)	0.03 (0.04)
Work–Life Climate (lower = better)	–0.01 (0.03)	0.02 (0.04)	–0.03 (0.04)

Notes: Remote effects remain close to zero across moderator levels; see Table 3 (Panels A–D) for main estimates.

6. Discussion

The rotating-exposure design across three firms shows that hybrid work consistently improves employee experience and modestly lifts output while fully remote work averages to near zero on productivity once role, firm, and calendar shocks are absorbed. These patterns align with the strongest available field evidence, which shows that randomized hybrid schedules raised satisfaction and reduced attrition without harming performance [1], and with network-level findings that fully remote

arrangements increase within-team closure and reduce cross-team bridging, which is often needed for complex work [2]. Identification relied on difference in differences estimators with event study diagnostics that address staggered timing and treatment heterogeneity [5], [6], [24], [65]. Pre trends were flat and alternative estimators reproduced the main pattern which supports the internal validity of the estimates [5], [6], [24], [65].

The study answered the research questions on average effects by showing that hybrid improved objective and subjective productivity, job satisfaction, and work–life climate, while remote did not reliably exceed on-site on those outcomes. The questions about mechanisms were addressed through mediation analyses and the integration of qualitative evidence, which indicated that autonomy and task–technology fit were the primary channels, with the communication structure playing a contextual role. The questions about boundary conditions were addressed by showing that higher role complexity attenuated gains, whereas stronger managerial support amplified them, aligning with the coordination and support literatures [19], [22], [23], [44]. The final question on robustness was addressed through event-study checks, heterogeneous robust estimators, and inference adjustments that left conclusions unchanged [5], [6], [24], [65]. In the same spirit, the hypotheses received consistent evidence. Hybrid outperformed on subjective productivity and satisfaction and delivered smaller yet positive movement in objective output, which supports H1a, H1b, and H2 in a manner that echoes the hybrid trial evidence [1] and classic work on telecommuting and satisfaction [34], [35]. Autonomy and task–technology fit explained a large share of the total effects, which supports H3, and the moderation by role complexity and managerial support supports H4 [20], [21], [14], [19], [22], [23], [44]. Where distance narrowed networks or video-constrained idea generation, remote did not surpass hybrid, which is consistent with recent evidence on collaboration structure and attention under mediated interaction [2], [42].

6.1 Theoretical implications

The findings connect work design with task–technology fit. Flexibility creates discretion and control over sequencing and timing, which lifts perceived efficiency and affect as suggested by autonomy research [20], [21]. The conversion of that potential energy into measurable output depended on the coordination architecture and the congruence between digital tools and task requirements, which is the core claim of task–technology fit [14]. Periodic co-presence in hybrid settings restored bridging ties and tacit alignment, which are hard to sustain at a distance, consistent with network evidence showing increased closure under fully remote conditions [2]. The results also help reconcile mixed conclusions in the macro literature by clarifying that effects are non-linear, with hybrid typically outperforming both ends of the spectrum when interdependence is non-trivial and when tool fit is adequate [27], [32]. Moderated mediation provides the mechanism sketch. Autonomy and fit are productive channels when coordination demands are bounded by cadence and managerial support, and they are weaker when complexity and interdependence are high and scaffolding is thin [19], [22], [23], [44]. The findings refine self-determination perspectives by showing that autonomy alone is insufficient without digital and coordination fit, extend task–technology fit by locating it within a broader motivational and coordination architecture, and qualify coordination theory by

documenting conditions under which hybrid designs can offset, rather than exacerbate, the coordination burdens of interdependence.

6.2 Managerial implications

Flexibility works when it is deliberately engineered rather than treated as a perk. A predictable cadence of in-person collaboration for interdependent work establishes shared context and reduces the coordination debt that accumulates in distributed settings, which keeps hybrid advantages durable [2]. Supportive management routines that specify weekly priorities, include short check-ins, and respect boundaries protect the work–life climate and amplify productivity gains, which align with evidence on perceived organizational support and family-supportive supervision [44], [19]. Tool integration matters because fragmentation in the collaboration stack raises search and switching costs while fit between workflow and digital platform allows autonomy to translate into throughput, which reflects task–technology fit [14]. For remote-first teams, success is most likely when tasks are modular and monitorable, and when feedback can be digitized at low cost, which echoes prior evidence on work-from-anywhere settings [3].

These recommendations travel most directly to large, knowledge-intensive multinationals with digital KPI and collaboration systems similar to those in our sample. For small and medium-sized enterprises, where HR analytics and formal performance dashboards are thinner, the same design principles can be implemented with simpler instruments: for example, anchoring hybrid around one or two fixed in-person days for the whole team, relying on a small set of readily observable output indicators, and using lightweight check-in routines rather than introducing new platforms. The core managerial tasks are to make interdependent work visible, agree ex ante on overlap hours, and explicitly protect boundaries so that flexibility does not slide into constant availability [27], [32], [44], [19]. In public sector organizations, where citizen-facing duties and statutory constraints limit remote work options, hybrid work is most feasible for back-office and analytical functions. There, fairness and transparency in eligibility criteria, rotation queues, and service coverage are central for maintaining legitimacy. Finally, in cultures or organizational climates with lower baseline autonomy and stronger norms of direct supervision, a gradual transition that couples modest schedule flexibility with supervisor training in outcome-based management is likely to be safer than abrupt shifts. In such settings, hybrid can be framed as a structured operating model rather than a discretionary perk, with explicit guardrails and feedback loops to surface problems early [27] – [29], [44], [19].

6.3 Limitations and future work

Observation covered six months, and longer horizons would allow study of learning accumulation network rewiring and promotion dynamics, which are often slower-moving outcomes [33]. Although the setting spans three multinationals and multiple role families, the institutional environment is Greek, a country with comparatively low baseline telework penetration and a specific regulatory framework for flexible work and digital work cards [29], [68], [57]. External validity is therefore strongest for knowledge-intensive organizations operating in similar low-to-medium telework, high-formality environments and weaker for sectors where tasks are less location-

independent, such as manufacturing, retail, or frontline public services [27], [32], [48], [49]. The participating firms are large and relatively digitally mature; small and medium-sized enterprises or public-sector organizations without comparable analytics and HR infrastructures are likely to adapt the rotation logic and measurement toolkit rather than replicate the design wholesale. Replication in high-WFH countries, in SMEs, and in public agencies would strengthen claims about transportability and clarify how institutional protections, collective bargaining, and sectoral constraints shape feasible hybrid models. Causal mediation with panels remains sensitive to timing and measurement choices, and we mitigated this by using sequential procedures and synchronizing qualitative evidence; however, experimental manipulations of cadence overlap hours and tool integration would provide sharper tests [5], [6], [24]. Finally, the paper reported standardized indices for comparability across roles, and we supplied mappings to natural units in the supplement to aid managerial interpretation while encouraging future work to extend those mappings with role-specific cost and value functions that capture the economics of output quality and cycle time [OS8].

7. Conclusions

This multi-firm rotating-exposure study shows that hybrid work yields modest gains in objective productivity and larger improvements in subjective productivity, job satisfaction, and work–life climate, whereas fully remote work averages near zero on productivity once role, firm, and calendar shocks are absorbed. These estimates draw on administrative performance indicators and validated surveys, analyzed using modern difference-in-differences and event-study diagnostics, and are reinforced by aligned qualitative evidence on day-to-day work patterns.

The contribution is an empirically grounded baseline for flexible work policy. By combining rotation across on-site, hybrid, and remote conditions, multi-method measurement, and identification strategies that address staggered timing and heterogeneous effects, the study provides credible, interpretable impacts that can be compared across roles and firms. The results indicate that well-governed hybrid arrangements can reconcile employee experience with organizational performance and offer a stable reference point for evaluating flexibility decisions.

Data Availability Statement: The data supporting the findings of this study are available from the corresponding author upon reasonable request. Supplementary material is available in the Online Supplement: Hybrid Work Design and Outcomes in Greek Multinationals.

Acknowledgements

This article received no financial or funding support.

Conflicts of Interest

The author confirms that there are no conflicts of interest.

References

[1] Bloom, N., Han, R. and Liang, J. Hybrid working from home improves retention without damaging performance. Nature, 2024, 630, 920–925.

[2] Yang, L., Holtz, D., Jaffe, S., Suri, S., Sinha, S. and Weston, J. The effects of remote work on collaboration among information workers. Nature Human Behaviour, 2022, 6, 43–54.

[3] Choudhury, P.R., Foroughi, C. and Larson, B. Work-from-anywhere: the productivity effects of geographic flexibility. Strategic Management Journal, 2021, 42(4), 655–683.

[4] Barrero, J.M., Bloom, N. and Davis, S.J. The evolution of work from home. Journal of Economic Perspectives, 2023, 37(4), 23–48.

[5] Sun, L. and Abraham, S. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. Journal of Econometrics, 2021, 225(2), 175–199.

[6] Goodman-Bacon, A. Difference-in-differences with variation in treatment timing. Journal of Econometrics, 2021, 225(2), 254–277.

[7] Kessler, R.C., Ames, M., Hymel, P.A., Loeppke, R., McKenas, D.K., Richling, D.E., Stang, P.E., Ustun, T.B. and Wang, P. The World Health Organization health and work performance questionnaire (HPQ). Journal of Occupational and Environmental Medicine, 2003, 45(2), 156–174.

[8] Koopmans, L., Bernaards, C.M., Hildebrandt, V.H., de Vet, H.C.W. and van der Beek, A.J. Construct validity of the individual work performance questionnaire. BMC Public Health, 2014, 14, 513.

[9] Bowling Green State University. Job descriptive index (JDI) and job in general. Bowling Green State University, 2020.

[10] Kinicki, A.J., McKee-Ryan, F.M., Schriesheim, C.A. and Carson, K. Assessing the construct validity of the job descriptive index: a review and meta-analysis. Journal of Applied Psychology, 2002, 87(1), 14–32.

[11] Tasios, T. and Giannouli, L. Job descriptive index (JDI): reliability and validity study in Greece. Psychology, Community and Health, 2017, 7(1).

[12] Sexton, J.B., Schwartz, S.P., Chadwick, W.A., Rehder, K.J., Bae, J., Bokovoy, J., Hunt, J.L. and Thomas, E.J. The associations between work–life balance behaviours, teamwork climate and safety climate: cross-sectional survey introducing the work–life climate scale. BMJ Quality and Safety, 2017, 26(8), 632–640.

[13] Schwartz, S.P., Adair, K.C., Bae, J., Rehder, K.J., Shanafelt, T.D., Profit, J., Sexton, J.B. and Thomas, E.J. Work–life balance behaviours cluster in work settings and relate to burnout and safety culture. BMJ Quality and Safety, 2019, 28(2), 142–150.

[14] Goodhue, D.L. and Thompson, R.L. Task–technology fit and individual performance. MIS Quarterly, 1995, 19(2), 213–236.

[15] Fetters, M.D., Curry, L.A. and Creswell, J.W. Achieving integration in mixed methods designs—principles and practices. Health Services Research, 2013, 48(6 Pt 2), 2134–2156.

[16] Guetterman, T.C., Fetters, M.D. and Creswell, J.W. Integrating quantitative and qualitative results in health science mixed methods research through joint displays. Annals of Family Medicine, 2015, 13(6), 554–561.

[17] Morgeson, F.P. and Humphrey, S.E. The work design questionnaire (WDQ): developing and validating a comprehensive measure for assessing job design and the nature of work. Journal of Applied Psychology, 2006, 91(6), 1321–1339.

[18] Eisenberger, R., Huntington, R., Hutchison, S. and Sowa, D. Perceived organizational support. Journal of Applied Psychology, 1986, 71(3), 500–507.

[19] Hammer, L.B., Kossek, E.E., Yragui, M.J., Bodner, T.E. and Hanson, G.C. Development and validation of a multidimensional measure of family-supportive supervisor behaviors (FSSB). Journal of Management, 2009, 35(4), 837–856.

[20] Humphrey, S.E., Nahrgang, J.D. and Morgeson, F.P. Integrating motivational, social, and contextual work design features: a meta-analytic summary and theoretical extension. Journal of Applied Psychology, 2007, 92(5), 1332–1356.

[21] Gagné, M. and Deci, E.L. Self-determination theory and work motivation. Journal of Organizational Behavior, 2005, 26(4), 331–362.

[22] Courtright, S.H., Thurgood, G.R. and Pierotti, A.J. Structural interdependence in teams: an integrative framework and meta-analysis. Journal of Applied Psychology, 2015, 100(6), 1825–1846.

[23] Okhuysen, G.A. and Bechky, B.A. Coordination in organizations: an integrative perspective. Academy of Management Annals, 2009, 3(1), 463–502.

[24] Callaway, B. and Sant'Anna, P.H.C. Difference-in-differences with multiple time periods. Journal of Econometrics, 2021, 225(2), 200–230.

[25] Eurofound. EWCS 2024: first findings—A new world of work: challenges and opportunities. Eurofound, 2024.

[26] Eurofound. The rise in telework: impact on working conditions and regulations. Eurofound, 2022.

[27] OECD. The surge of teleworking: a new tool for local development? OECD, 2023.

[28] Aksoy, C.G., Barrero, J.M., Bloom, N., Davis, S.J., Dolls, M. and Zarate, P. Working from home around the world. Brookings Papers on Economic Activity, 2022.

[29] EURES (EU). Labor market information—Greece: remote work. EURES, 2025.

[30] Bloom, N., Liang, J., Roberts, J. and Ying, Z.J. Does working from home work? evidence from a Chinese experiment. NBER Working Paper No. 18871, 2013.

[31] Gibbs, M., Mengel, F. and Siemroth, C. Work from home and productivity: evidence from personnel and analytics data on IT professionals. Journal of Political Economy Microeconomics, 2023, 1(1), 7–41.

[32] OECD. Compendium of productivity indicators 2023. OECD, 2023.

[33] Emanuel, N., Harrington, E. and Pallais, A. The power of proximity to coworkers: training for tomorrow or productivity today? NBER Working Paper No. 31880, 2023.

[34] Gajendran, R.S. and Harrison, D.A. The good, the bad, and the unknown about telecommuting: meta-analysis of psychological mediators and individual consequences. Journal of Applied Psychology, 2007, 92(6), 1524–1541.

[35] Allen, T.D., Golden, T.D. and Shockley, K.M. How effective is telecommuting? assessing the status of our scientific findings. Psychological Science in the Public Interest, 2015, 16(2), 40–68.

[36] Lyzwinski, L.N., Foroughi, S.R., Booth, C.S. and Smith, M.A. A scoping review of remote work and health. International Journal of Environmental Research and Public Health, 2024.

[37] DORA Research Program. 2018 state of DevOps report. DORA, 2018.

[38] Kaplan, R.S. and Norton, D.P. The balanced scorecard—Measures that drive performance. Harvard Business Review, 1992.

[39] Kessler, R.C. Using the WHO health and work performance questionnaire (HPQ) to evaluate the indirect workplace costs of illness. Harvard Medical School, 2004.

[40] Podsakoff, P.M., MacKenzie, S.B., Lee, J.-Y. and Podsakoff, N.P. Common method biases in behavioral research: a critical review of the literature and recommended remedies. Journal of Applied Psychology, 2003, 88(5), 879–903.

[41] Slemp, G.R., Kern, J.R., Vella-Brodrick, C. and Martin, R.J. Leader autonomy support in the workplace: a meta-analytic review. Psychological Bulletin, 2018, 144(12), 1205–1227.

[42] Brucks, M.S. and Levav, J. Virtual communication curbs creative idea generation. Nature, 2022, 605, 108–112.

[43] Leonardi, P.M. When flexible routines meet flexible technologies: affordance, constraint, and the imbrication of human and material agencies. MIS Quarterly, 2011, 35(1), 147–167.

[44] Kurtessis, J.N., Eisenberger, R., Ford, M.T., Buffardi, L.C., Stewart, K.A. and Adis, C.S. Perceived organizational support: a meta-analytic evaluation of organizational support theory. Journal of Management, 2017, 43(6), 1854–1884.

[45] Hussey, M.A. and Hughes, J.P. Design and analysis of stepped wedge cluster randomized trials. Contemporary Clinical Trials, 2007, 28, 182–191.

[46] Hemming, K., Haines, T.P., Chilton, P.J., Girling, A.J. and Lilford, R.J. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. BMJ, 2015, 350, h391.

[47] Senn, S.J. Cross-over trials in clinical research. Chichester, U.K.: Wiley, 2002.

[48] Deaton, A. and Cartwright, N. Understanding and misunderstanding randomized controlled trials. Social Science and Medicine, 2018, 210, 2–21.

[49] Shadish, W.R., Cook, T.D. and Campbell, D.T. Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton Mifflin, 2002.

[50] Hayes, A.F. Introduction to mediation, moderation, and conditional process analysis: a regression-based approach. New York, NY: Guilford, 2018.

[51] Pearl, J. Causality: models, reasoning, and inference. Cambridge, U.K.: Cambridge University Press, 2009.

[52] Barrero, J.M., Bloom, N. and Davis, S.J. Why working from home will stick. NBER Working Paper No. 28731, 2021.

[53] Lim, C.Y. et al. Considerations for crossover design in clinical study. Contemporary Clinical Trials Communications, 2021, 23.

[54] Dwan, K. et al. Reporting of randomised crossover trials: CONSORT extension. BMJ, 2019, 366, l4378.

[55] Hemming, K., Taljaard, M. and Grimshaw, J. Introducing the new CONSORT extension for stepped-wedge cluster randomised trials. Trials, 2019, 20, 68.

[56] Wasserman, S. and Faust, K. Social network analysis: methods and applications. Cambridge, U.K.: Cambridge University Press, 1994.

[57] EY Greece. Work time schedule and digital work card—obligations and ERGANI II integration. EY Greece, 2022.

[58] Cameron, A.C. and Miller, D.L. A practitioner's guide to cluster-robust inference. Journal of Human Resources, 2015, 50(2), 317–372.

[59] Braun, V. and Clarke, V. Using thematic analysis in psychology. Qualitative Research in Psychology, 2006, 3(2), 77–101.

[60] Gioia, D.A., Corley, K.G. and Hamilton, A.L. Seeking qualitative rigor in inductive research: notes on the Gioia methodology. Organizational Research Methods, 2013, 16(1), 15–31.

[61] Krueger, R.A. and Casey, M.A. Focus groups: a practical guide for applied research. Thousand Oaks, CA: Sage,

2015.

[62] Krippendorff, K. Content analysis: an introduction to its methodology. Thousand Oaks, CA: Sage, 2019.

[63] Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20(1), 37–46.

[64] ALLEA. European code of conduct for research integrity. ALLEA, 2023.

[65] Pustejovsky, J.E. and Tipton, E. Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. Journal of Business and Economic Statistics, 2018, 36(4), 672–683.

[66] Rambachan, A. and Roth, J. A more credible approach to parallel trends. Econometrica, 2023, 91(6), 2283–2309.

[67] Maxwell, J.A. The importance of qualitative research for causal explanation. Qualitative Inquiry, 2012, 18(8), 655–661.

[68] Hellenic Ministry of Labour and Social Affairs. Νόμος 4808/2021—ρυθμίσεις για σύγχρονες μορφές εργασίας (art. 67: telework). Government Gazette A’101, 2021.