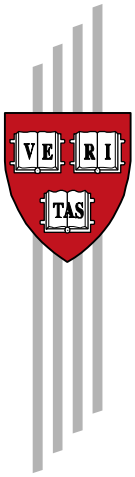
Policy Decisions and Evidence Use among Civil Servants: A Group Decision Experiment in Pakistan

Policy Decisions and Evidence Use among Civil Servants: A Group Decision Experiment in Pakistan

Laura Metzger, Theodore Svoronos, and Adnan Qadir Khan

CID Faculty Working Paper No. 377
April 2020



Working Papers

Center for International Development at Harvard University

Policy decisions and evidence use among civil servants. A group decision experiment in Pakistan

Laura Metzger[a,*], Teddy Svoronos[b], Adnan Qadir Khan[c]

Abstract:

In a lab-in-field experiment with elite civil servants in Pakistan, we investigate whether groups outperform individuals in a two-staged task which requires effective use of data and evidence. We also study how efficiently groups harness their members' individual knowledge for problem-solving. We do not find a significant difference in individual (first stage) and group performance (second stage). Yet, groups could have significantly improved their performance during the second stage of the task, had they more efficiently collaborated to retrieve their members' respective knowledge. Carefully interpreted in the setting of our experiment, our data suggests that diversity in individual knowledge may hamper effective use of data and evidence for decision-making in small groups of policymakers.

Key Words: evidence-based policy, adult learning, group decisions, lab-in-field experiment, civil servants, Pakistan
JEL classification: A29, C92, D37, D38, I28, I38
Affiliated Research Program: Evidence for Policy Design (EPoD)

[a]John F. Kennedy School of Government, Harvard University, United States; * Corresponding author: laura_metzger@hks.harvard.edu
[b] John F. Kennedy School of Government, Harvard University, United States
[c] School of Public Policy, London School of Economics, United Kingdom

Acknowledgements: We thank our colleagues at the Kennedy School's Evidence for Policy Design (EPoD) and the Center for Economic Research in Pakistan (CERP) for supporting our research. We are particularly grateful for the expertise and assistance provided by Amelia Knudson, Emily Myers, Charlotte Tuminelli, Amna Aaqil, Jehanara Amin and Ghania Suhail. We are indebted to the Civil Services Academy (CSA) in Lahore for their collaboration on this project and for invaluable support during its implementation. Amelia Knudson and Laura Metzger further want to thank Zulfiqar Younas (Director of the Common Training Program at CSA) and his staff for their exceptional hospitality.

1. Building capacity for evidence-based policy

1.1 Background

Researchers and practitioners from different disciplines increasingly call for more evidence-based public policy. This demand is motivated by the desire to spend resources on policies that have a measurable welfare impact and, thus, address social problems effectively. The underlying assumption is that data and evidence will help policymakers decide what interventions are best to invest in. Clearly, this is easier said than done: Policymakers act in complex environments that may or may not favor the use of data and evidence, and not all policy domains lend themselves equally well to data collection and evidence use (Gugerty and Karlan 2018). Even under favorable conditions, some key factors are required for making evidence-based policy feasible. Technical skills that enable policymakers to use data and evidence are one such factor. This research is embedded in one of the largest capacity building efforts to date that tackles technical barriers to evidence use in government, the *Building Capacity to Use Research* Evidence (BCURE) initiative. The Harvard BCURE initiative was housed at Harvard Kennedy School's Evidence for Policy Design (EPoD) program. Since BCURE's 2013-2017 run and under

new funding since, EPoD has trained over 4,500 civil servants in India, Pakistan, Nepal, and Bangladesh on the use of data and evidence for decision-making. Thus, BCURE is a highly welfare-relevant, at-scale initiative with the potential to affect the lives of hundreds of thousands of citizens through the training it provides to government officers.[1]

1.2 Research Objectives

How effectively do trainings like BCURE strengthen evidence-based policy? How do policymakers engage with evidence for decision-making, and how does evidence use affect decision-making quality? Recent studies in development economics and public policy provide first insights into important questions like these (Baekgaard, Christensen, Dahlmann, Mathiasen, and Petersen, 2017; Banuri, Dercon, and Gauri 2018; Coville and Vivalt, 2019; Hjort, Moreira, Rao, and Santini, 2019). Using a lab-in-field experiment, we explore two novel questions in this thematic area. One, do civil servants working on a policy problem in small groups outperform officers working on the same problem individually? Two, how efficiently do officers learn from each other and benefit from each other's knowledge by collaborating (peer learning)? Group decisions are an extremely relevant study topic considering that all kinds of choices – from private household consumption to national public policies – are often made jointly with others (by families, professional teams etc.). Moreover, professional organizations increasingly entrust teams with important decisions, signaling that group decisions provide value added over individual decisions (Charness and Sutter, 2012). Even where decisions are not made by formally established groups, we are safe to assume that in their working contexts, policymakers are likely to deliberate, discuss and seek approval from others when taking decisions.

Within the scope of group decisions, we are interested in two aspects of the relationship between evidence use, decision making, and decision quality. The first is whether group work can boost learning outcomes among civil servants in the context of technical training. Second, we want to gain a first understanding of whether group work can be a useful tool to foster evidence-based decision-making and be applicable to officers' working contexts. We are not aware of other quantitative studies investigating the impact of capacity building training on decision quality among policymakers. We are also not aware of other lab-in-field experiments studying group decisions in a sample of policymakers. *Figure 1* below depicts the five main literature strands that provide the broader analytical framework for this exploratory research, and to which we contribute. The literature review in the next section presents studies that are immediately relevant to our research.

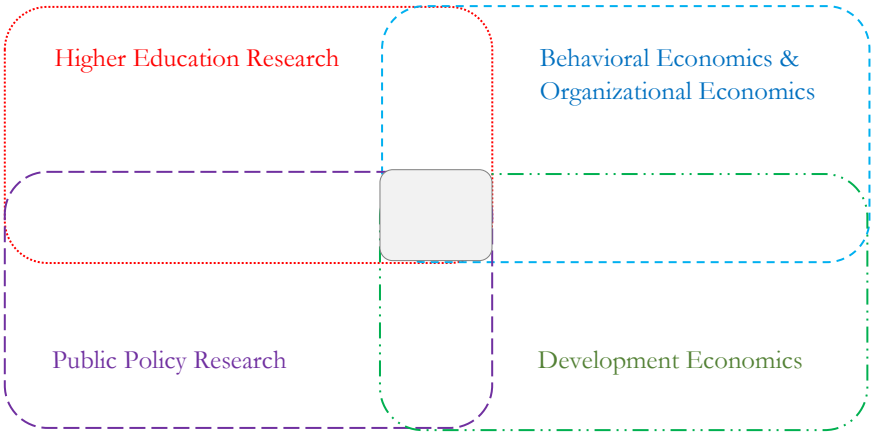


Figure 1: Embedding in the literature at the intersection of five disciplines

[1] We define "policymakers" broadly and include civil servants, politicians, and other government staff. We use the terms policymakers, civil servants and (government) officers interchangeably throughout the paper.

2. Learning and decision making in groups

2.1 Evidence-based decision-making

Over the past 20 years, social science research has generated a wealth of evidence on the welfare impact of policy interventions in countries around the globe. New data collection technologies, methodological progress in impact measurement, and stronger emphasis on applied research have helped generate this body of evidence. Development economics and public policy research have been actively contributing to pushing this work forward (Banerjee and Duflo 2011, Banerjee and Duflo 2019). Yet, in what *way* policymakers use evidence for decision-making is unclear. Several recently published studies address this knowledge gap (see Hjort, Moreira, Rao, and Santini 2019; Kalla and Porter 2019; Coville and Vivalt 2019; Banuri, Dercon, and Gauri 2018; Baekgaard, Christensen, Dahlmann, Mathiasen, and Petersen 2017). Most of this new research focuses on the prevalence of behavioral biases among individual policymakers which, if present, may lead to sub-optimal decision outcomes and welfare losses. While decision biases of private individuals are well-documented in the literature (Simon 1955; Schelling 1968; Kahnemann and Tversky 1979), much less is known about policymakers. Considering that policymakers are experts with a mandate to make decisions in the best interest of society, one might expect that their decision-making is less prone to "typical" biases (Coville and Vivalt 2019).

With this hypothesis in mind, Banuri, Dercon, and Gauri (2018) conduct a survey experiment on a sample of full-time staff of UK's Department for International Development (DFID) and the World Bank to test for loss aversion[2] (Tversky and Kahnemann 1981) and confirmation bias.[3] Loss aversion describes a different behavioral response to the same decision problem when it is framed as a loss instead of a gain; confirmation bias is used in the psychological literature to describe the tendency to seek and interpret information in a way that aligns with existing beliefs and expectations. The study finds evidence for both, loss aversion as well as confirmation bias. A follow-up experiment with a small sample of DFID staff suggests that group discussions can mitigate confirmation bias but not loss aversion. The authors suggest this might be the case because the confirmation bias experiment involved recognizing the right answer which, apparently, is facilitated through group deliberation. However, this does not apply to the loss aversion experiment. Vivalt and Coville (2019) also run a survey experiment on policymakers and researchers who are employed by or otherwise related to DFID or the World Bank to investigate behavioral biases in this sample. The study presents individuals with real research evidence on programs that seek to increase school enrollment in developing countries. It finds evidence for optimism bias and for "variance neglect". Optimism bias occurs when individuals update more strongly on good news than bad news, and variance neglect means that policymakers do not sufficiently consider the dispersion of potential welfare impacts when assessing an intervention. However, the study also finds that providing more information fosters rational belief updating. Vivalt and Coville (2019) replicate the experiment on Amazon's Mechanical Turk with average citizens but fail to find a difference between this sample and the policy experts.

Baekgaard, Christensen, Dahlmann, Mathiasen, and Petersen (2017) test whether politicians systematically interpret evidence in a way that is consistent with their ideological leaning. This process is referred to in the psychological literature as motivated reasoning (see also Kahan, Peters, Dawson, and Slovic 2017). The study is based on a randomized survey experiment with Danish politicians, which Baekgaard et al. repeat with Danish citizens to investigate differences between both population groups. The data suggests that politicians and citizens are biased by their prior political leaning when interpreting evidence, and that adding more (contradictory) evidence does not seem to lead to rational updating. This contrasts the findings by Coville and Vivalt (2019) that more information does lead to more rational updating. However, Baekgaard et al. (2017)

[2] Usually, people are risk-seeking in losses and risk-averse in gains. However, it has been shown that this type of bias is mitigated by presenting information in a more accessible way.

[3] The term confirmation bias is used in the psychological literature to describe the tendency to seek and interpret information in a way that aligns with existing beliefs, expectations or hypothesis. Cognitive psychologist Peter Wason originally coined the term.

present their sample with fictitious information about topics that were politically loaded in Denmark at the time (childcare and elderly care), while Coville and Vivalt (2019) present real evidence on a more neutral topic (child education). The political saliency of a topic may increase individuals' propensity to motivated reasoning (see Bénabou, 2015) and could be one reason for diverging results in both studies. Besides that, Danish citizens and policymakers may have taken the fictitious information less seriously.

Hjort, Moreira, Rao, and Santini (2019) conduct two field experiments with 657 municipal leaders, mostly mayors, from Brazil. Their first experiment investigates (a) whether leaders demand research evidence to learn about the policy impact of early childhood education programs (by eliciting their willingness to pay for evidence) and (b) whether they update their prior beliefs about program impacts after receiving new evidence. Like the above research, this study also tests for deviations from Bayesian learning, namely confirmation bias, optimism bias, and motivated reasoning. Hjort et al. find that policymakers *do* update their beliefs based on information provided to them, and do *not* find evidence for confirmation bias, optimism bias, or motivated reasoning. In a second experiment, the study explores whether leaders are willing to adopt new policies based on evidence provided to them. A subset of their sample received detailed information about interventions that are proven to effectively raise taxes. Results from a follow-up survey, conducted between 15 to 24 months later after the second experiment, show that policymakers did act on the received information: The chance of policy adoption in municipalities increased by 10 percentage points when municipal leaders were provided with evidence about a successful policy intervention. An important feature of this study is that policymakers were faced with evidence highly relevant to their contexts - early childhood education (something Brazil's central government had already been pushing for) and raising taxes. Thus, policymakers' willingness to update their beliefs and adopt new successful policies based on proven results is potentially higher in this set up than in related studies. In sum, existing research on the relationship between evidence-use and decision-making among policymakers shows mixed results. However, it documents that decision biases exist in this population group.

Our research contributes to this new evidence base, further contributing to behavioral social sciences and higher education research. In contrast to previous studies, we focus on a new topic (group decisions) and a previously unstudied sample (bureaucrats in a lower income country).

Although our experiment is not designed to address a specific question in organizational economics, it relates to this field of research as well. By focusing on group decisions among policymakers we ultimately strive to open new lines of inquiry about decision-making, organizational structure and optimal acquisition and use of information in government organizations (Gibbon and Roberts 2013, Milgrom and Roberts 1992).

2.2 Group decisions and peer learning

Economic theory predicts that individual and group decisions should not differ under complete information and rational decision-making (Kocher and Sutter 2005), but empirical studies provide evidence to the contrary (see e.g., Janis 1972; Hart 1994; Stasser, Taylor, and Hanna 1995; Blinder and Morgan 2005; Blinder and Morgan 2007; Kocher and Sutter 2005; Sutter 2010; Kugler, Kausel, and Kocher 2012). Several studies have shown that groups outperform individuals at problem solving, recalling information, mitigating decision-making biases, and pooling intellectual resources for problem solving (Blinder and Morgan 2005; Kocher and Sutter 2005; Laughlin, Hatch, Silver, Boh 2006; Blinder and Morgan 2008; Sutter 2010; Kugler, Kausel, and Kocher 2012; Carey and Laughlin 2012; Banuri, Dercon, and Gauri 2018; Levy, Svoronos, and Klinger 2018). Groups can also perform worse than individuals, however. For instance, by polarizing opinions, groupthink (a desire for harmony and conformity), and by failing to efficiently exploit the knowledge of individual group members (Janis 1972; Stasser and Steward 1992; Stasser and Steward 1995; Kugler, Kausel, and Kocher 2012; Hastie and Sunstein 2015). Although behavioral research on group decisions is comprehensive (see Charness and Sutter 2012, for a discussion of behavioral economics research), important gaps remain. This includes the study of group decision-making in government and the ways in which such group decisions affect public policy design.

Why are group decisions particularly interesting in a public policy context? Like other organizations, public administrations delegate tasks and responsibilities to teams. It is important to understand how this affects the decisions civil servants take on behalf of and for citizens. Group work may serve as a tool to improve the quality of decision-making. To the extent that groups outperform individuals in solving tasks that require evaluating data and quantitative evidence, resource-constrained administrations in lower income countries may see efficiency and quality gains by delegating such tasks to groups. Indeed, "a lack of manpower and funds to collect and analyze data or to conduct and interpret research" is considered a significant barrier by civil servants in our sample (Harvard 2015, Harvard 2018). Group-decision research is typically based on laboratory experiments with student samples in high income countries, which are very different from the high-powered civil servants the Harvard BCURE program works with. Thus, a first area to explore is whether the finding that groups can make better decisions than individuals (Blinder and Morgan, 2005; Blinder and Morgan, 2007; Kocher and Sutter 2005; Lombardelli, Proudman and Talbot, 2005; Banuri, Dercon, and Gauri, 2018) generalizes to civil servants as well. A training intervention we conducted with senior administrative service officers in India indicates that groups seem to perform better than individuals when it comes to correctly evaluating and interpreting research evidence on a policy problem. Second, groups may help foster peer learning and evidence use in the classroom and, thus, help overcome technical barriers to using data and evidence effectively. Indeed, experimental research by Lombardelli, Proudman and Talbot (2005) suggests that individuals learn from each other's decisions and can fare better in groups than individually; yet, they also find that groups are not significantly better than their most able member.

Another strand of literature providing important insights into peer learning and group decisions is higher education research. Studies by Lasry, Mazur and Watkins (2008), Jang, Lasry, Miller and Masur (2017) as well as Levy, Svoronos and Klinger[4] (2018) show positive effects of peer instruction and collaborative learning on students' learning and motivation. Levy, Svoronos and Klinger (2018) introduce an indicator to quantify peer learning. Collaborative efficiency, as they call it, describes the efficiency with which groups retrieve knowledge that is available to them through their individual members. The indicator also captures whether groups create new knowledge when deliberating on a problem (see Section 3.5 for a detailed description). Between 2013 and 2017, Levy, Svoronos and Klinger (2018) collected data on 900 Harvard students during statistics and econometrics courses that are part of the school's Masters programs in International Development and Public Policy. Students in these courses take two-stage exams. The first stage is an individual closed-booked examination. In the second stage, students are randomly assigned to groups of 3 to 5 individuals and must solve the most challenging subset of stage 1 questions collaboratively. Leaning on this design, we employ a two-stage task in our experiment as well (see Section 3.3). The findings of Levy, Svoronos and Klinger (2018) indicate that groups significantly improve their performance when moving from the first to the second stage of the exam. This in turn suggests that groups benefit from their members' skill and knowledge. At the same time, the average collaborative efficiency was 0.68. This means that students retrieved 68% of the knowledge available in their group, which indicates that there is considerable room for exploiting benefits from peer learning.

3. Setting and Experimental Design

3.1 Local setting

We conducted our experiment during two days of BCURE training at the Civil Services Academy[5] (CSA) in Lahore, Pakistan, in November 2019.[6] The CSA was established in 1948 as a training academy for fresh entrants to the Pakistan Administrative Services. It is one of the most prestigious academies in the country providing general training to civil servants to prepare them for their professional career. Each new cohort begins their training in September and officers reside on campus for the entire eight months of training. The curriculum

[4] See Levy, Svoronos, and Klinger (2018) for a more in-depth discussion of research on peer learning and two-stage examinations.
[5] http://csa.edu.pk/ (last accessed March 31, 2020)
[6] The experiment is approved by Harvard's Internal Review Board (IRB).

consists of lectures, seminars, workshops, co-curricular activities and field visits. All activities are mandatory. Attendance is handled strictly and, since 2019, monitored with biometric fingerprint scanners. A typical day starts with physical exercise at 6 am followed by curricular and co-curricular activities and ends at around 4 pm. Officers are free to leave campus during weekends or during the week after class.

BCURE training has been embedded in the mandatory CSA curriculum since 2017. BCURE content consists of six thematic modules[7] taught with blended learning techniques. Blended learning combines online training and in-class sessions and allows instructors to see how learners responded in online training, so that they can customize in-class content to learners' needs. Evidence - for the most part concentrated on North American colleges and universities - suggests that blended learning can lead to more acquisition of skills and higher student performance than face-to-face instruction (Bazelais and Doleck 2018).[8] Each module consists of a 90 minute[9] online training on one day, followed by 90-minute in-class sessions (lectures and case exercises) the next day. Thus, in total, CSA officers spend about 18 hours on BCURE training. Performance is graded and counts 4% towards an officer's final graduation grade at CSA. Modules are taught by local faculty who have been trained by Harvard faculty during a Training of Trainers event at Harvard, as well as by Harvard faculty or Harvard faculty affiliates who travel to the country for this specific purpose.

We chose two of the six training modules: 'Aggregating Evidence' and 'Impact Evaluation' for our experiment. Both modules focus on using data and evidence for policy decisions and, thus, incorporate the very essence of the evidence-based policy debate. All material we used in the experiment – the training protocol for research assistants, the instructions and the task – is provided in Appendix E. For the purposes of this experiment, we leveraged content alignment between the modules to combine two 90-minute in-class sessions into one 180-minute session. We implemented the experiment during this 180-minute session (see Section 3.3 for details). The modules were taught by Adnan Qadir Khan, Harvard affiliate and co-author on this paper.

3.2 Sample

The current cohort, the 47th common training program, counts 270 officers who started their training at CSA in September 2019. In order to matriculate into the common training program at CSA (and the civil service itself) officers must first have passed a competitive national exam, the Central Superior Service (CSS) exam. The exam is administered by the Federal Public Service Commission to recruit candidates for Federal Government services, including, for example, Commerce and Trade, Pakistan Administrative Service, Inland Revenue Service, Foreign Services, and the Police Services of Pakistan.[10] Applicants must have a Bachelor's degree with at least second class division (out of a total of three divisions). Admission rates to the civil services are low; around 3% or less. In 2019, 14,521 applicants presented themselves for the written exam. Hence, those working for the Federal Government are a group of highly select individuals, many of whom occupy influential government positions over the course of their career. Given permissible age limits, fresh entrants are between 18 and 32 years old.

We recruited the entire cohort of 270 officers for our experiment. Due to sickness related absences or no-shows, our final sample consists of 261 individuals. Female officers make up 39% of the sample. As a region,

[7] The modules are: Systematic Approaches to Policy Design; Descriptive Evidence; Aggregating Evidence; Impact Evaluations; Cost-Benefit Analyses; Commissioning Evidence; Officers complete all modules during training. A seventh module, Using Data Systems, was created in 2018 and is not yet embedded in CSA training.

[8] It is important to emphasize that more rigorous evidence on blended learning is needed to unveil the efficacy of this new pedagogical approach more broadly. In the context of BCURE, blended learning is a beneficial training tool given that servants have less study time available than regular university students since they receive these trainings while already being on the job. The online modules, to which they are granted full access, allow them to (re)engage with the content outside of the classroom as well and at their convenience.

[9] Officers are free to use more or less than 90 minutes to complete the online part.

[10] Details about the exam, and a complete list of the Federal Government services officers are recruited for can be found here: https://www.studyandexam.com/css-exam-general-information.html (last accessed March 31, 2020)

Punjab is overrepresented, with 123 officers originating from there. Details on officers' regions of origin, and their assignment to different service groups are provided in Appendix A.

3.3 Control and treatment group set-up

This experiment has three key features.

1. Officers were randomly assigned to work on a task individually (control group) or in groups of two (treatment group).
2. We administered a two-stage task in which officers solved the same set of problems twice: first during stage 1 and a second time during stage 2. (More information about the task is presented in section 3.4, below.)

Officers in the treatment group solved the task individually in stage 1 and switched to group work in stage 2. Groups were randomly assigned. Officers in the control group solved the task individually during stage 1 and stage 2. The group versus individual work during stage 2 was the *only* difference between the treatment and control groups. *Figure 2* depicts the experimental design and shows that it results in a difference-in-difference framework. This framework allows us to (a) analyze the effect of group work on peer learning and decision quality (i.e., task performance), and (b) distinguish this effect from the effect that learning over time may have on decision quality. Learning over time can occur irrespective of group work given that officers engage with the same task twice.

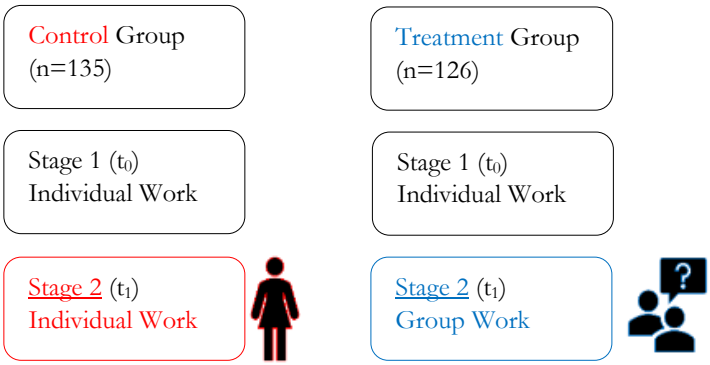


Figure 2: Main experimental design

In the control group, communication with others was prohibited during both stages. In the treatment group, communicating with others was prohibited during stage 1. During stage 2, group members could communicate freely within their group but were not allowed to communicate with other groups. Research assistants closely monitored the following of this rule. Moreover, group members had to agree on and submit a joint answer to each task problem. Varying the rules under which group members agree on an answer to see how it effects decision outcomes was not a priority, since groups consisted of two members only. Varying group size was not a priority either, given sample size limitations. Power calculations, based on data we collected during a pilot intervention with senior administrative officers in India in June 2019, suggested that a group size of two would be appropriate. With 126 individual observations in the treatment group, we were left with 63 group observations during stage 2.

We fully integrated the experimental design with the two BCURE training modules 'Aggregating Evidence' and 'Impact Evaluation'. This means that both the lecture and task were designed to reinforce the concepts in the BCURE online learning modules, which officers had completed during the previous two days. Moreover, the task challenged officers to actively use content for problem-solving. As mentioned earlier, we implemented the experiment during a 180-minute in-class session. The in-class session consisted of two core components: a

lecture and the experimental task (see section 3.4). The first part of the lecture prepared officers for the task with a review of key concepts and warm-up round with easy questions that could be answered by thoroughly reading the task text. Officers then completed the task. After the task was completed, we moved to the last part of the lecture which was devoted to giving officers immediate feedback on their performance and guiding a discussion of the correct answers to the task questions. Providing instant feedback allowed us to discuss officer's perception of the task and the group work. It also made the officers more active and informed participants in the experiment, making the exercise more interesting overall. *Figure 3* illustrates the sequencing of these activities.

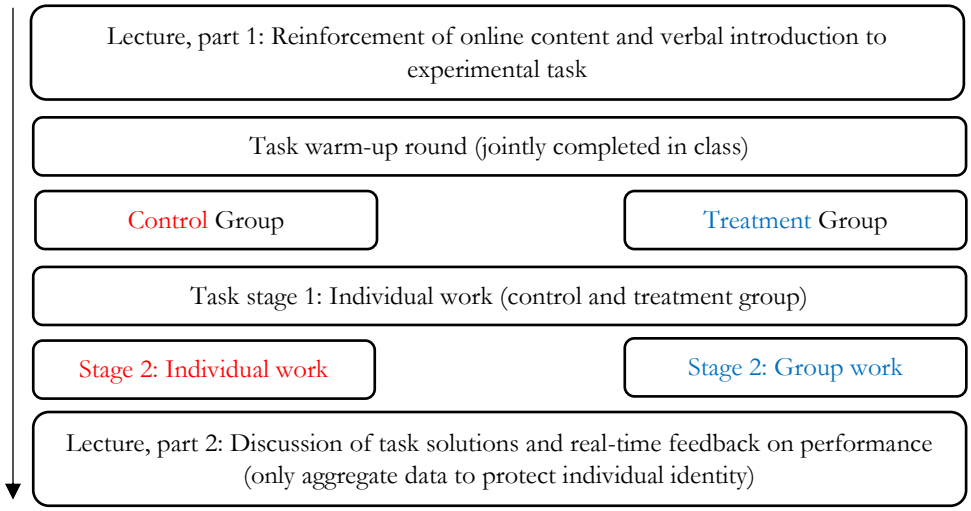


Figure 3: Time sequence of actions

3.4 Task and task incentivization

We designed a text-based task that required officers to answer a set of graded, multiple-choice questions with only a single correct option. The actual influence of the grade on individuals' overall final grade was negligible which officers were aware of. However, individual performance is closely monitored by CSA faculty and is taken very seriously by the officers. Thus, grading the task provided a strong incentive to perform well. Each correctly answered question was assigned one point; wrongly answered questions were assigned zero points. The maximum possible was score was 10. The task consisted of assessing three pieces of existing evidence (a randomized control trial, a matching study, and a difference-in-difference study) to decide whether a planned drinking water intervention was likely to be effective in reducing waterborne diseases in children under five and increasing household income. In sum, the task consisted of assessing a realistic policy problem with data and research evidence.

Prior to moving into stage 1, officers acquainted themselves with the task during a warm-up round that was jointly completed in-class. By a show of hands, officers voted on the correct answers to six multiple-choice questions about the task that were projected on the lecture slides. The correct answer to each warm-up question was immediately discussed in class. The warm-up questions did not overlap with the questions encountered by officers during stages 1 and 2. The purpose of the warm-up round was to make sure that everyone entered stage 1 with the same basic knowledge about the task's content; it did not affect the grading in any way.

For both treatment and control, Stage 2 involved answering the same 10 multiple choice questions a second time (either individually or in groups). During this time they had the opportunity to review and adjust their stage 1 answers and improve their score. During stage 1, officers were given 30 minutes to answer as many

questions as possible. For stage 2 they were given 20 minutes. In between stages, we scheduled a 15-minute break. We reduced the time available for solving the task during stage 2 to make it sufficiently challenging for the officers.

For logistical reasons, the intervention was administered to two groups on two separate days, with a total of 161 officers participating during day 1 and a total of 109 officers participating during day 2. To avoid spillovers between days, we altered the task questions such that the correct answers were different on both days, but with no loss to comparability. All task-related material, including the instructions, is presented in Appendix E.

3.5 Variables

To reiterate, our main research questions are:

1. Do groups perform better than individuals at problem-solving and, hence, make better quality decisions?

   Our indicator for decision quality in the control and treatment group is task performance. This is our first main dependent variable. Since we keep the task constant throughout, we can compare task performance between control and treatment group as well as between task stages. Task performance is defined as the sum of earned points divided by the maximum possible points (=10). For example, an officer who scores 4 points during stage 1 would have a performance-on-task score of 0.4. Two officers in the same group scoring 3 points during stage 2 would have a group score of 0.3 points.

2. How efficient is peer learning?

   Our indicator for peer learning is collaborative efficiency. This is our second main dependent variable and it is specific to the treatment group. Collaborative efficiency is defined as the group score divided by the "super student score", which represents the combined knowledge that individual group members bring to the group stage. For example, imagine that the task consists of a total of three questions, where each question has one right answer. Thus, 3 is the maximum score. During stage 1, group member 1 scores 1 point on question one, 0 points on question two, and 1 point on question three. Group member 2 scores 0 points on question one, 1 point on question two and 1 point on the question three. In this case, the combined knowledge of both group members, the super student score, is: 1 point (question one) + 1 point (question two) + 1 point (question three) = 3 points. This is because although they only scored 2 points each, each group member answered a different set of questions correctly during stage 1. Thus, each member knows one item the other does not know. Figure 4 illustrates the super student score. Further imagine that the stage 2 group score is equal to 2. This group's collaborative efficiency will be: 2 (group score)/3 (super student score) = 0.67. This means that the group retrieved 67% of the knowledge that was theoretically available to it when entering the group stage. Correspondingly, a group with a collaborative efficiency score of 1 retrieved 100% of the knowledge available to it. Scores above 1 mean that groups created new knowledge by collaborating. *Figure 4* visualizes the calculation of the collaborative efficiency indicator.

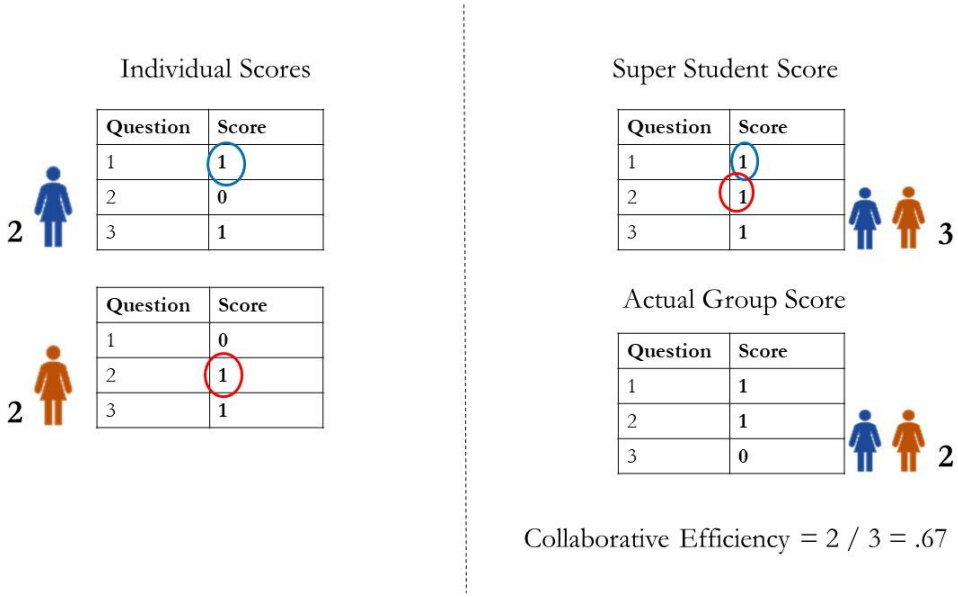


Figure 4: Calculation collaborative efficiency = Group Score/ Super Student Score. Knowledge that is specific to individual group members is circled in blue and red.

Further important variables we use are the individual average score, the top student score, the top surplus, and the super surplus.

- The individual average score is the average of the stage 1 scores for officers in the same group. It allows us to compare performance across both stages in the treatment group.

  In the above example, the individual average score for each group member would amount to (2+2)/2 = 2.

- The top student score is the highest score achieved by an officer in his or her group.

  In the above example the top student score is 2 and it happens to be the same for both group members.

- The top surplus is the difference between the top student score and the individual average score. It provides us with a measure of how far apart two officers are in terms of their score.

  In the above example, the difference is 0 points.

- The super surplus is the difference between the super student score and the individual average score. It provides us with a measure of the diversity of knowledge in a group.

  In the above example, the difference is 3 – 2 = 1.

Table 1 lists the formal definitions of all variables we use in our data analysis (adapted to our context from Levy, Svoronos, and Klinger, 2018).

Table 1: Variable definitions

Variable name	Definition	
Raw Scores		
Individual score#	$p = X_{ij1} = \sum_{k=1}^{m} Q_{kijs}$	Stage 1 score
Group score	$gp_j = X_{ij2}$	Stage 2 score
Dependent variables		
Performance on task as proportion (score)	$p = {}^{1}/_{10} \sum_{k=1}^{m} X_{ijs}$	Stage 1 and Stage 2 score
Performance gain	gp_j - IAS_j	Stage 2 – Stage 1 score
Collaborative efficiency	group score/ super student score	Stage 2 score
Variables for control and treatment group analysis		
Treatment group dummy	treatment group=1; control group=0	
Stage 2 dummy	task stage 2=1; task stage 1=0	
Day 2 dummy	day 2 session=1; day 1 session=0	
Female officer dummy	female=1; male=0	
Officer's region of origin (entering the analysis as separate dummies)	1=Punjab; 2= Khyber Pakhtunkhwa; 3=Sindh Urban; 4=Balochistan; 5=Sindh Rural; 6=Azad Jammu and Kashmir; 7=Federally Administered Tribal Areas	
Variables for treatment group analysis		
Individual average score	$IAS_j = 1/n_j \sum_{i=1}^{n_j} X_{ij1}$	Based on stage 1 scores
Top student score	$TSS_j = MAX_{i=1,\dots n_j} X_{ij1}$	
Super student score	$SSS_j = \sum_{k=1}^{m} MAX_{i=1,\dots n_j} (X_{ij1})$	
Gain	gp_j - IAS_j	Differences between Stage 2 and Stage 1 scores
Top Surplus	$TSS_j - IAS_j$	
Super Surplus	$SSS_j - IAS_j$	

Notation: n_j is group size (=2 in our case); m is the number of questions (=10 in our case); k is the question; i is the student; j is the group; s is the stage.

The score obtained in question k by student i in stage s is expressed as Q_{kijs}

3.6 Data collection and instant feedback

As explained previously, officers received instant feedback about their performance during the last part of the lecture. Since the task was paper-based, we needed a technology to digitalize and process answers on the spot. We used Zip-Grade[11] to prepare answer sheets that are scannable and that can be graded instantly using a smart phone. We scanned the sheets right after stage 2, exported the data in csv file format and processed it in STATA

[11] https://www.zipgrade.com/ (last accessed March 31, 2020)

to produce statistics that revealed the (difference) in average performance between control and treatment group, as well as peer learning in the treatment group.

4. Results

4.1 Main effects

4.1.1 Control group

The control group allows us to test whether learning did occur between Stage 1 and Stage 2 considering that officers repeat the task during the second stage and, thus, are given more time to work on the same set of questions. Learning over time may occur independently of peer learning, and we want to distinguish between the two effects. Appendix B presents statistical analysis showing that learning over time did *not* occur in the control group. Most officers did not change their answer at all between stages (Figure B1 and Figure B2). We also observe that officers performed better on the second day overall (Table B1 and B2), and that female officers performed significantly better than male officers (Table B3).

4.1.2 Control and treatment group: Difference-in-Difference

We estimate the following regression to evaluate the treatment effect:

$$Y_i = \alpha + \beta T_i + \gamma t_i + \delta (T_i \cdot t_i) + \varepsilon_i$$

The dependent variable, Y_i, represents performance on task expressed as a proportion (see *Table* 1); α represents the constant; β represents the difference between control and treatment groups during the first stage; γ represents learning between stages in the control group. The DiD estimator, δ, indicates the differential effect of group work on task performance; i.e., the estimated difference between the change in performance in the control group (Stage 2 – Stage 1) and the change in performance in the treatment group (Stage 2 – Stage 1).

The analysis presented in *Table 2* reveals that we do not find a statistically significant difference between treatment and control group, or between stage 1 and stage 2 regarding task performance. We also do not observe a significant differential effect of group work on task performance. In other words, we do not observe positive effects of group work or learning over time on decision quality. That groups do not perform better than individuals contrast the findings from the July 2019 pilot study we conducted with Indian civil servants, as well as the positive learning effects Levy, Svoronos, and Klinger (2018) document for their collaborative two-stage exams at Harvard.

In Appendix C, we provide summary statistics of the distribution of earned points in control and treatment group. It is noteworthy that, on average, officers answered 4.2 out of 10 questions correctly. The fact that officers answered less than 50% of the questions correctly suggests that the task was sufficiently, maybe even excessively, challenging. This is also borne out by previous work that suggests that there are serious constraints to policymakers' ability to interpret evidence (Callen et al. 2017).

Table 2: OLS regressions, joint analysis for control and treatment group

Total of earned points expressed as proportion	Main effects	Session effects	Main effects; Session effects	Controls	DiD	DiD; Session effects	DiD; Session effects; Controls
Treatment Group (β)	0.0114 [0.0185]		0.0115 [0.0184]		0.00730 [0.0236]	0.0104 [0.0256]	0.0111 [0.0234]
Stage 2 (γ)	0.0143 [0.0192]		0.0143 [0.0191]		0.0104 [0.0259]	0.00746 [0.0235]	0.0104 [0.0250]
Treat.##Stage 2 (DiD)					0.0103 [0.0381]	0.0103 [0.0380]	0.0109 [0.0382]
Day 2		0.0494* [0.0192]	0.0494* [0.0192]			0.0494* [0.0192]	0.0478* [0.0191]
Female Officer				0.0484* [0.0195]			0.0470* [0.0191]
Officer's region of origin (Punjab omitted)				Yes			Yes
Constant	0.411*** [0.0159]	0.403*** [0.0113]	0.392*** [0.0168]	0.418*** [0.0157]	0.413*** [0.0180]	0.394*** [0.0187]	0.390*** [0.0208]
Observations	459	459	459	459	459	459	459
R-squared	0.002	0.015	0.017	0.030	0.002	0.017	0.047

White-Huber robust standard errors in brackets; * p<0.05; ** p<0.01; *** p<0.001

4.1.3 Treatment group analysis

Table 3 provides summary statistics of the variables that we use to analyze decision quality (task performance) and peer learning in the treatment group (see *Table 1*, Section 3.5 for variable definitions). *Figure 5* depicts the distribution of the individual average score, the group score, the super student score, and the top student score. The mass of observations for the super student score is located to the right of the individual average score. The same applies to the top student score, but the difference is less pronounced. Two-sided t-tests confirm that the difference in means between the individual average score and the super student score, as well as between the individual score and the top student score are statistically significant at 1 percent at least.[12] This suggests that groups have room to improve their stage 1 scores during stage 2, either by following the lead of the top student, or by effectively harnessing their pooled knowledge. The data also show that groups can benefit more by collaborating effectively than following the top student. *Figure 6* visualizes the distribution of collaborative efficiency scores. On average, groups retrieved about 70% of the knowledge that was theoretically available to them through their members. Few groups created new knowledge during the second stage.

[12] SSS (mean 6.33) = IAS (mean 4.21), t=17.85, ***p<0.001; TSS (mean 5.08) = IAS (mean 4.21), t=9.16, ***p<0.001

Table 3: Summary statistics treatment group, both intervention days (observations pooled)

Variable	Mean	Standard Deviation
Individual Average Score (IAS) (Stage 1)	**4.21**	1.27
Super Student Score (SSS) (Stage 1)	**6.33**	1.53
Top Student Score (TSS) (Stage 1)	**5.08**	1.53
Super Surplus (Stage 1)	2.13	0.95
Top Surplus (Stage 1)	0.87	0.76
Gain (Stage 2 – Stage 1)	0.21	1.42
Group Score (Stage 2)	4.41	1.87
Collaborative Efficiency Score (Stage 2)	0.72	0.28
Observations (both intervention days)[1]	63	

[1]See Appendix 3 for a split of these summary statistics by intervention day

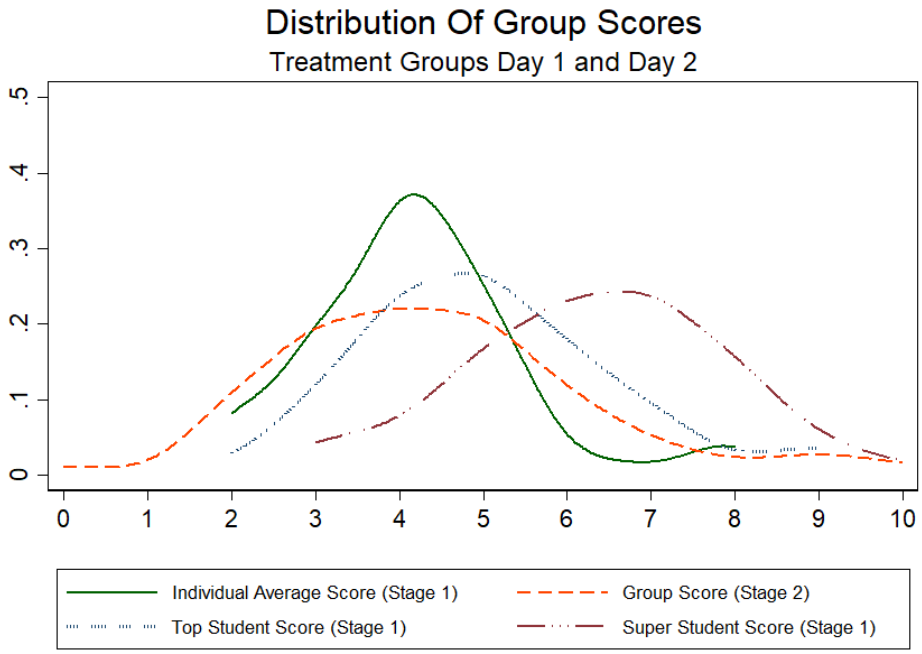


Figure 5: Treatment group scores, observations pooled across both intervention days

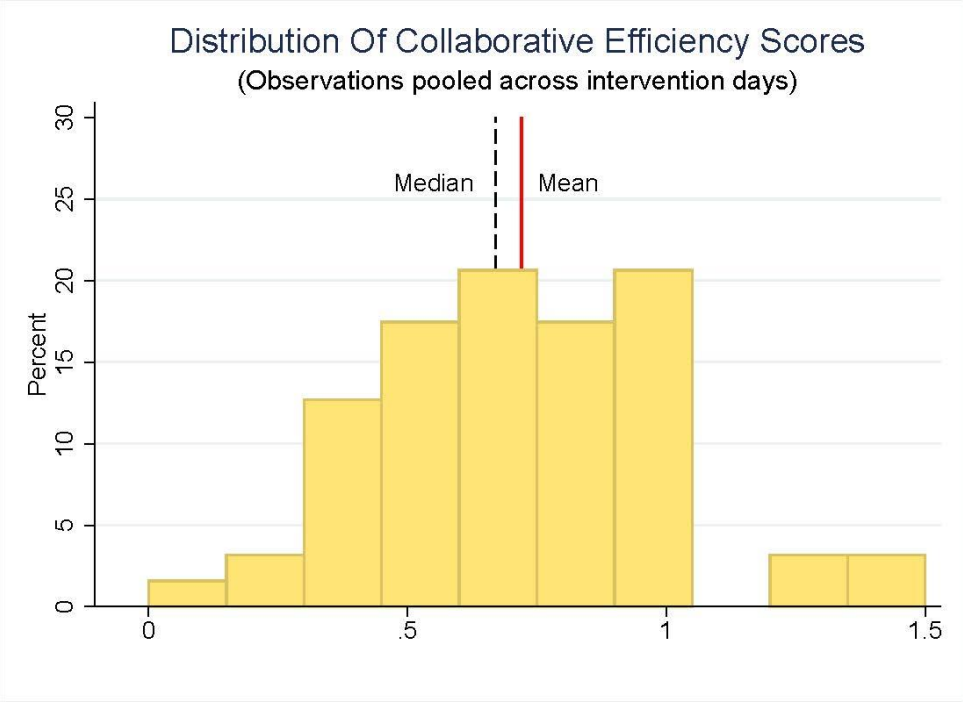


Figure 6: Collaborative efficiency scores, pooled across both intervention days

Table 4 presents analysis of the potential drivers of task performance (left side of the table) and collaborative efficiency (right side of the table) in the treatment group. The two standout findings are, first, that a larger difference between the better ("top") and the weaker officer in a group has a positive effect on performance gains. Second, a larger difference in individual knowledge between two group members, indicated by a larger super surplus, has a significant negative effect on collaborative efficiency. In other words, group members are less likely to improve their score, the more diverse the knowledge in their group; they are more likely to improve their score when a stronger officer is paired up with a weaker officer. In contrast, Levy, Svoronos and Klinger (2018) find that the super surplus has a significant positive effect on performance gains and collaborative efficiency in their Harvard student sample.

One possible interpretation of the first finding is simply that when individual abilities and/or expertise levels are farther apart, group decisions provide a greater margin of superiority over individual decisions (Blinder and Morgan 2008). From an organizational performance perspective, this could imply that delegating certain tasks to groups composed of individuals with diverse expertise and abilities may indeed lead to better decision outcomes than letting staff work on such tasks individually. Regarding the second finding, it may be possible that individual group members who possess similar expertise and/or abilities in the same domain (here: interpreting data and evidence on public policies) but know different things well within that domain have more difficulties agreeing on an answer, with negative effects on overall decision quality. These findings raise a number of interesting questions for future research that we further discuss below (e.g. optimal group composition and size).

Table 4: OLS regression, treatment group, correlates of performance gains and collaborative efficiency (all variables are standardized)

	Performance gains (Stage 2 score - Stage 1 score)						Collaborative efficiency					
Individual Average Score (Stage 1 group average)	-0.0329 [0.1325]				-0.0793 [0.1240]	-0.0824 [0.1276]	0.0463 [0.1348]				-0.0166 [0.1018]	-0.0173 [0.1045]
Top Surplus		0.1675 [0.1459]			0.3251* [0.1457]	0.3226* [0.1460]		-0.0361 [0. 1324]			0.2769* [0.1132]	0.2760* [0.1118]
Super Surplus			-0.1700 [0.1175]		-0.3195** [0.0972]	-0.3133** [0.1053]			-0.5486*** [0.0989]		-0.6876*** [0.0931]	-0.6749*** [0.1012]
Day 2				0.1094 [0.2529]	0.0317 [0.2543]	0.0338 [0.2573]				0.3289 [0.2432]	-0.0542 [0.2263]	-0.0440 [0.2270]
Male group (vs. mixed)						0.0187 [0.2745]						-0.0395 [0.2316]
Female group (vs. mixed)						0.0538 [0.3212]						0.0498 [0.2668]
Officer's region of origin (Punjab omitted category)						Yes						Yes
Constant	0.0000 [0.000]	0.0000 [0.253]	0.0000 [0.1251]	-0.0434 [0.1706]	-0.0126 [0.1593]	-0.0322 [0.2038]	-0.0000 [0.1268]	-0.0000 [0.1269]	-0.0000 [0. 1061]	-0.1305 [0.1760]	0.0215 [0.1395]	0.0242 [0.1983]
Observations	63	63	63	63	63	63	63	63	63	63	63	63
R-squared	0.001	0.003	0.029	0.003	0.112	0.112	0.0021	0.0013	0.3010	0.0263	0. 3621	0.3631

White-Huber robust standard errors in brackets; * p<0.05; ** p<0.01; *** p<0.001

4.2 Carryover effects

Although we do not find that group work significantly affected performance in the treatment group, we tested whether the experience of group work affected officers' performance in posttests that the BCURE program conducts to test knowledge gains acquired through training. We do not find that treatment group participants perform significantly better than control group participants on post-tests for either the Impact Evaluation or Aggregating Evidence modules (see Table C2, Appendix C).

5. Discussion and conclusion

We implemented a lab-in-field experiment to explore potential benefits of group decisions on learning outcomes and decision-making quality in a sample of elite civil servants in Pakistan. Our research is embedded in one of the largest existing training initiatives that aims to strengthen data and evidence use among policymakers. The first research question we wanted to answer is whether civil servants working on a policy problem in small groups outperform civil servants working on the same problem individually. The second question is how efficiently officers learn from each other and benefit from peer learning. Insights generated from this research can help us understand whether group work could be an effective tool to boost learning outcomes among civil servants in the context of technical training. This is important given that without the necessary skills in place moving towards evidence-driven public policy is hard to achieve. After all, insufficient technical skills are considered a top barrier to data and evidence use in government. To the extent that group work positively affects training and learning outcomes, we may ask whether it can also be a useful tool to foster evidence-based decision-making in real organizational contexts.

In contrast to the positive learning outcomes observed in previous studies (Levy, Svoronos, and Klinger 2018; Lasry, Mazur, and Watkins 2008), we do not find a significant positive effect of two-stage exams or group work on learning outcomes of CTP officers. However, we do find that by more efficiently pooling and harnessing their individual knowledge, groups could have significantly improved their decision quality. Interestingly, we find that if the difference in skills (for the specific task in this intervention) between two group members is larger, the group score seems to be driven by the top student, which in turn results in a better group performance overall. When the individual knowledge of two group members is rather diverse (fewer overlap between correctly answered questions), it seems to be more difficult for them to agree on an answer which, in turn, results in a lower group performance. This is a cautious interpretation of our results in the context of our specific experiment and may, of course, change if the setting was different. Also, it is worth emphasizing that since this is the first study of this kind, its results do not imply that group work should be discarded either as a pedagogical tool or an effective way to improve decision-making in an organizational context. More research is necessary to explore the manifold layers of group decisions.

In that sense, our findings do raise several follow-up questions such as how collaborative efficiency could be increased, about optimal group size and composition, or how group decisions would play out in an organizational context. For example, a larger group size might help mitigate disagreements over the correct answers to test questions and, in addition, increase problem solving capacity. On the other hand, a larger group size can also exacerbate the problem of coordination and free-riding and, thus, incentives to contribute. As a study by Laughlin, Hatch, Silver, Jonathan, and Lee (2006) suggests, groups with three to five members (compared to 2-person groups) significantly increase the chance of finding a correct solution to "highly intellective" problems. Another intervention could be to provide guiding questions for groups, or some other form of decision aids, to help them solve the problem at hand and focus on relevant informational clues. This may increase the demonstrability of a problem, i.e. group members' ability to still recognize a correct solution to a problem even if they initially failed to solve it. Demonstrability, in turn, can help increase group performance (Amir, Amir, Shahar, Hart, and Gal 2018). The interplay between group size *and* demonstrability

and its effect on performance is a further interesting topic to explore and not yet too well understood (ibid). Another implication of our study is to highlight the importance of incentives to acquire and use knowledge dispersed through the organization as well as the appropriate balance between specialization and knowledge transfer (Lazear and Gibbs 2009).

Last, we want to discuss our findings in light of critical issues related to our experimental design and the local context. First, we think that the difficulty level of the task may have been rather challenging since officers answered less than 50% of the questions correctly, on average. The concepts we asked about went, to some extent, beyond the standard content of the BCURE modules. Thus, the demonstrability of the experimental task might have been too low. Consequently, group members who did not correctly answer the task questions may have been less likely to recognize solutions proposed by those who did. More extensive pre-testing might have helped to better calibrate a more appropriate difficulty level and is something worth considering for future interventions. This is especially true since the educational background and career specialization of officers, despite their generally high skill level, is rather heterogeneous. Second, based on side conversations with officers, the intervention "felt" rather formal (due to, e.g., sealed envelopes and a highly organized schedule) and created a test-like atmosphere that may have put subjective pressure on officers to perform. The fact that the lecture and intervention were implemented by external faculty and researchers may have added to this atmosphere. Third, we communicated openly about the fact that officers are participating in an experiment. This in turn may have distracted officers from focusing on the task in a more relaxed manner. Fourth, conversations with CSA faculty revealed that officers are not used to collaborating with each other. These attitudes may be driven by the fact that this is the first stage of training for these fresh entrants into civil service after they have been selected through a highly competitive process where allocation to services is determined on the basis of their ranks. As an aside, since these officers are expected to work in teams through their careers this raises the question for CSA of leveraging the initial socialization to generate an appropriate willingness to work in teams. At Harvard, group work is an integral component of the learning environment. This applies to other Western student samples most studies are relying on as well. This difference might provide one explanation why our findings on collaborative learning and group work differ from previous work such as the study by Levy, Svoronos, and Klinger (2018). Finally, we initially raised the question whether group decisions could potentially promote data- and evidence-driven decision making in an organizational context. The lessons we can extract from this study now with regards to organizational settings are limited. But our experiment is a first important step towards building up relevant knowledge. Conducting this research in a controlled environment with a relevant sample allows us to identify "basic" mechanisms that drive group decisions and evidence use. As we better understand those mechanisms, we can incrementally increase the layer of complexity that comes with examining decision-making in organizational contexts.

A final implication of our work is to highlight that it is not only important focusing on what evidence and knowledge policymakers need to make better policies but also to examine how policymakers actually understand, learn and interpret such evidence. This also feeds into the question of devoting more resources and time to learning about the demand for new evidence and about the barriers that make it hard for policymakers to learn and use that evidence (Velasco 2019).

References

Amir Ofra, Amir Dor, Shahar Yuval, Hart Yuval, and Gal Kobi. 2018. "The more the merrier? Increasing group size may be detrimental to decision-making performance in nominal groups." *PLoS ONE* 13(2): e0192213

Banerjee Abijit and Duflo Esther. 2011. "Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty", First edition, Public Affairs.

Banerjee Abijit and Duflo Esther. 2019. "Good Economics for Hard Times", First edition, Public Affairs.

Bénabou Roland. 2015. "The Economics of Motivated Beliefs," Revue d'économie politique, Dalloz, Vol. 125(5), pp. 665-685

Bazelais Paul and Doleck Tenzin. 2018. "Investigating the impact of blended learning on academic performance in a first semester college physics course", *Journal of Computers in Education* **5**, 67-96(2018)

Baekgaard Martin, Christensen Julian, Dahlmann Casper M., Mathiasen Asbjørn and Petersen Grund Niels B. 2017. "The role of evidence in politics: Motivated reasoning and persuasion among politicians." *British Journal of Political Science* 08, 1–24.

Michael Callen, **Adnan Khan**, **Asim I. Khwaja**, **Asad Liaqat** and **Emily Myers**. 2017. "These 3 barriers make it hard for policymakers to use the evidence that development researchers produce. Monkey Cage. *Washington Post*.

Charness Gary and Sutter Matthias. 2012. "Groups Make Better Self-Interested Decisions." *Journal of Economic Perspectives* 26 (3): 157-76.

Sheheryar Banuri, Stefan Dercon and Varun Gauri. 2019. "Biased Policy Professionals" *The World Bank Economic Review* Volume 33, Issue 2, pp. 310–327

Blinder Alan and Morgan John. 2005. "Are Two Heads Better than One? Monetary Policy by Committee" *Journal of Money, Credit and Banking* Vol. 37, No. 5, pp. 789-811

Blinder Alan and Morgan John. 2008. "Leadership in Groups: A Monetary Policy Experiment" *International Journal of Central Banking* Vol. 4(4), pp. 117-150

Carey Harold R. and Laughlin Patrick. 2012. Groups perform better than the best individuals on letters-to-numbers problems: Effects of induced strategies. *Group Processes & Intergroup Relations* 15(2), 231–242

t'Hart Paul. 1994. "Groupthink in government: a study of small groups and policy failure" Baltimore: John Hopkins University Press

Gibbons Robert and Roberts John. 2013. *The Handbook of Organizational Economics*. Princeton University Press

Gugerty Mary K. and Karlan Dean. 2018. "Ten Reasons Not to Measure Impact – and What to Do Instead.", Stanford Social Innovation Review, Summer 2018

Harvard Kennedy School (HKS), Evidence for Policy Design (EPoD). 2015. "BCURE Training Needs Assessment Report", January 15/ 2015, Harvard Kennedy School, Cambridge, Massachusetts, unpublished.

Harvard Kennedy School (HKS), Evidence for Policy Design (EPoD). 2018. "Training Assessment and Follow-Up Support for LBSNAA Training.", March 30/ 2018, Harvard Kennedy School, Cambridge, Massachusetts, unpublished.

Hastie Reid and Sunstein Cass R. 2014. "Making Dumb Groups Smarter." Harvard Business Review, December 2014

Hjort Jonas, Moreira Diana, Rao Gautham and Santini Juan F., 2019. "How Research Affects Policy: Experimental Evidence from 2,150 Brazilian Municipalities" NBER Working Paper No. w25941

Jang Hyewon, Lasry Nathaniel, Miller Kelly and Mazur Eric. 2017. "Collaborative exams: Cheating? Or learning?" *American Journal of Physics* 85, 223

Janis Irving L. 1972. "Victims of Groupthink." New York: Houghton Mifflin.

Kahan Dan M., Peters Ellen, Dawson Erica C., Slovic Paul. 2017. "Motivated numeracy and enlightened self-government." *Behavioral Public Policy* Volume 1, Issue 1, 54-86

Kahneman Daniel and Tversky Amon. 1979."Prospect theory: An analysis of decision under risk." Econometrica 47(2), 263–291

Kalla Joshua and Porter Ethan. 2019. "Correcting Bias in Perceptions of Public Opinion Among American Elected Officials: Results from Two Field Experiments." *British Journal of Political Science*, forthcoming

Kugler Tamar, Kausel Edgar E. and Kocher Martin G. 2012. "Are groups more rational than individuals? A review of interactive decision making in groups." *Advanced Review* Volume 3, July/August 2012

Lasry Nathaniel, Mazur Eric and Watkins Jessica. 2008. *American Journal of Physics* 76, 1066

Laughlin Patrick, Hatch Erin, Silver Jonathan, and Boh Lee. 2006. Groups perform better than the best individuals on letter-to-numbers problems: effects of group size. *Journal of Personality and Social Psychology* 4 (4):644–651

Lazear Edward and Gibbs Michael. 2009. Personnel Economics in Practice. John Wiley & Sons, Inc.

Levy Dan, Svoronos Theodore, Klinger Mae. 2018. "Two stage examinations: Can examinations be more formative experiences?" *Active Learning in Higher Education*, pp. 1-16

Milgrom Paul and Roberts John. 1992. *Economics, Organizations and Management*. Prentice Hall, Englewood Cliffs, NJ 07632.

Schelling Thomas. 1968. "Problems in public expenditure analysis" Washington, DC: Brookings Institute., chapter: The life you save may be your own., pp. 127–176

Simon Herbert A. 1955. "A Behavioral Model of Rational Choice" *The Quarterly Journal of Economics* Vol. 69, No.1, pp. 99-118.

Stasser Garold and Stewart Dennis. 1992. "Discovery of Hidden Profiles by Decision-Making Groups: Solving a Problem Versus Making a Judgement." *Journal of Personality and Social Psychology* Vol. 63., No. 3, 426-434

Stasser Garold and Stewart Dennis. 1995. "Expert Roles and Information Exchange during Discussion: The Importance of Knowing Who Knows What." *Journal of Experimental Social Psychology* 31, 244-265

Sutter Matthias. 2005. "Are four heads better than two? An experimental beauty-contest game with teams of different size." *Economic Letters* 88(2005) 41-46

Vivalt Eva and Coville Aidan. 2019. "How do policymakers update?"

Andrés Velasco. 2019. "Bipolar Economics." Blog. Project Syndicate. https://www.project-syndicate.org/commentary/limits-of-randomized-controlled-economics-trials-by-andres-velasco-2019-11?barrier=accesspaylog (last accessed March 31, 2020)

Appendix A – Summary Statistics

Variable	Percentage share	Observations (control and treatment group)
Female officers	.39	261
Breakdown by officer's region of origin	100.00	261
Punjab	47.15	123
Khyber Pakhtunkhwa	13.03	34
Sindh Urban	14.56	38
Balochistan	4.06	12
Sindh Rural	12.64	33
Azad Jammu and Kasmir	1.92	5
Federally Administered Tribal Areas	5.75	15
Breakdown by officer's service group assignment	100.00	261
Commerce and Trade Group	7.66	20
Foreign Service of Pakistan	7.58	19
Information Group	2.3	6
Inland Revenue Services	14.56	38
Military Land and Cantonment Group	2.3	6
Office Management Group	28.74	75
Pakistan Audit and Account Services	8.05	21
Pakistan Administrative Services	14.56	38
Pakistan Customs Service Group	7.28	19
Postal Group	1.91	5
Police Service of Pakistan	4.98	13
Railway Commercial and Transport Group	0.38	1

Appendix B – Control group analysis

Table B1: Summary statistics of score gains & earned points in the control group

Variable	Mean	Standard Deviation	Observations
Day 1 and Day 2 observations pooled			
Gain (Stage 2 score – Stage 1 score)	0.11	1.04	135
Total earned points†, Stage 1	4.13	2.09	135
Total earned points†, Stage 2	4.24	2.16	135
Day 1 observations only			
Gain (Stage 2 – Stage 1)	0.11	1.08	81
Total earned points†, Stage 1	3.86	1.92	81
Total earned points†, Stage 2	3.97	2.01	81
Day 2 observations only			
Gain (Stage 2 – Stage 1)	0.09	.98	54
Total earned points†, Stage 1	4.56	2.28	54
Total earned points†, Stage 2	4.65	2.34	54

† Maximum scorable points = 10

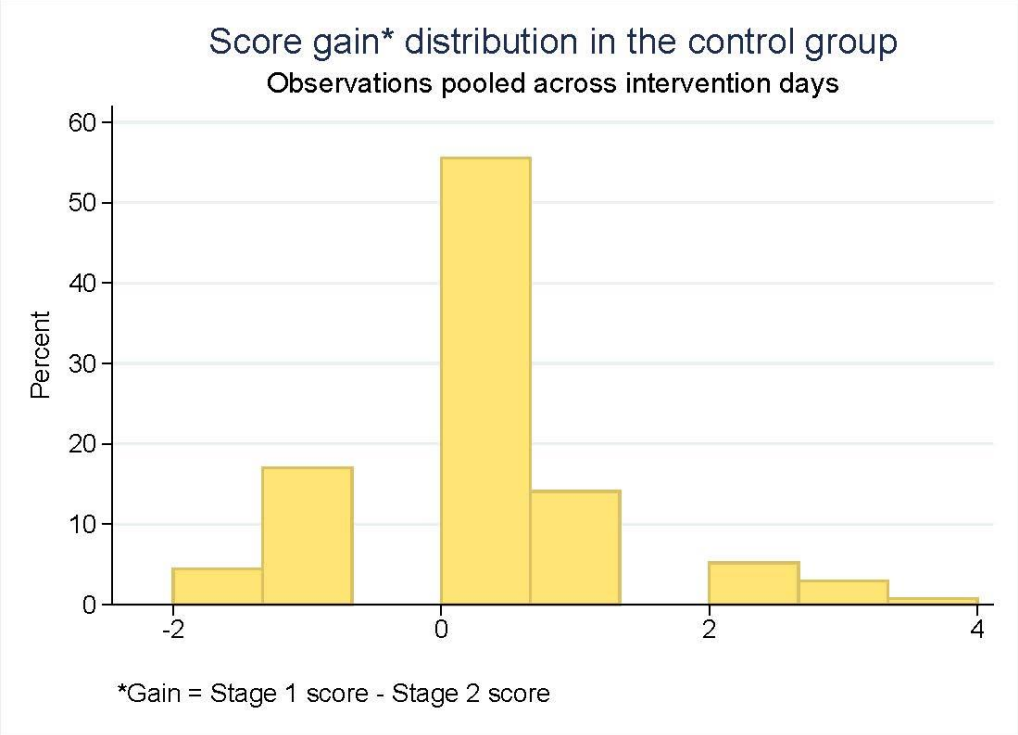


Figure B1: Distribution of score gains in percent, control group

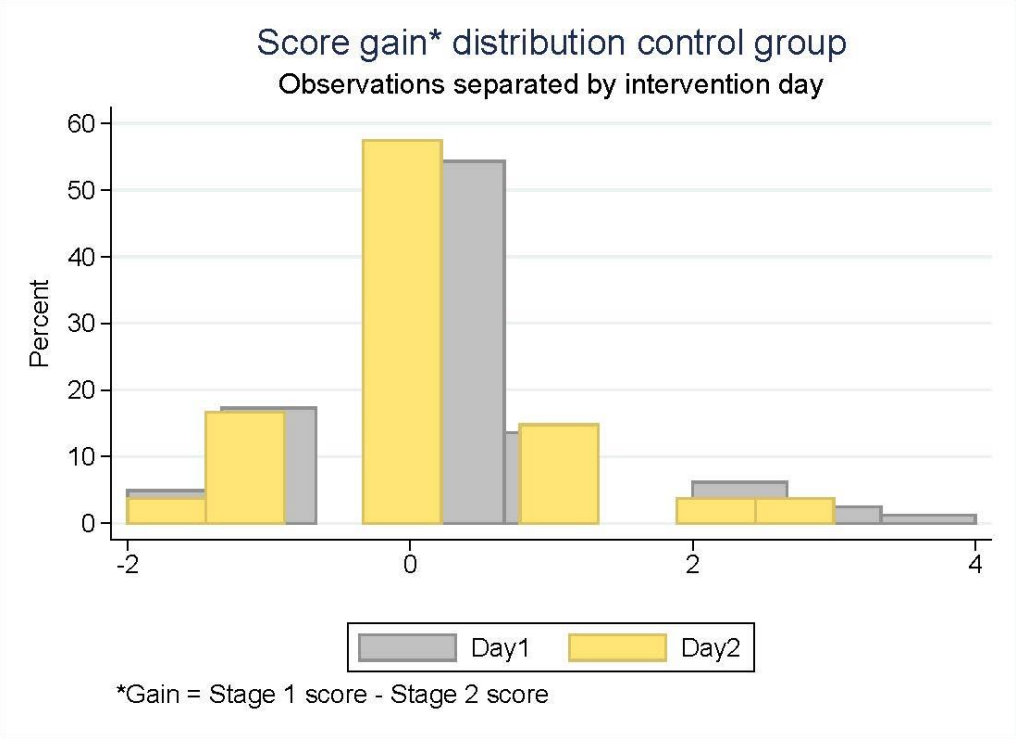


Figure B2: Distribution of score gains in percent by intervention day, control group

Table B3: Testing for session effect

Variable	Mean difference between day 1 and day 2	T-statistic	Observations
Gain (Stage 2 score – Stage 1 score)	0.02	0.10	135
Total earned points†, Stage 1	-0.71	-1.94[1]	135
Total earned points†, Stage 2	-0.69	-1.82[1]	135

* p<0.05; ** p<0.01; *** p<0.001 ([1]significant at 10% at least)

† Maximum scorable points = 10

Table B2: OLS regression, control group

Total of earned points expressed as proportion	Stage & Session Effects	Stage & Session Effects, Controls
Stage 2	0.0104 [0.0256]	0.0104 0.0248
Day 2	0.0694* [0.0269]	0.0680** [0.0254]
Female Officer		0.0632* [0.0236]
Officer's region of origin (Punjab omitted category)	Yes	Yes
Constant	0.386*** [0.0197]	0.382*** [0.0236]
Observations	270	270
R-squared	0.027	0.11

* p<0.05; ** p<0.01; *** p<0.001 – White-Huber robust standard errors in brackets

Appendix C – Control and treatment group analysis

Table C1: Summary statistics of earned points in the control and treatment group; observations pooled across intervention days

Variable	Mean	Standard Deviation	Observations
Stage 1 and Stage 2 observations pooled			
Total of earned points†	4.22	1.97	459
Total of earned points in percent‡	.42	.97	459
Stage 1			
Total earned points†, Stage 1	4.17	1.91	261
Total earned points in percent‡, Stage 2	.42	.19	261
Stage 2			
Total earned points†, Stage 1	4.29	2.1	198
Total earned points in percent‡, Stage 2	.43	.21	198

†Maximum scorable points = 10; ‡Scored points/10 max. scorable points

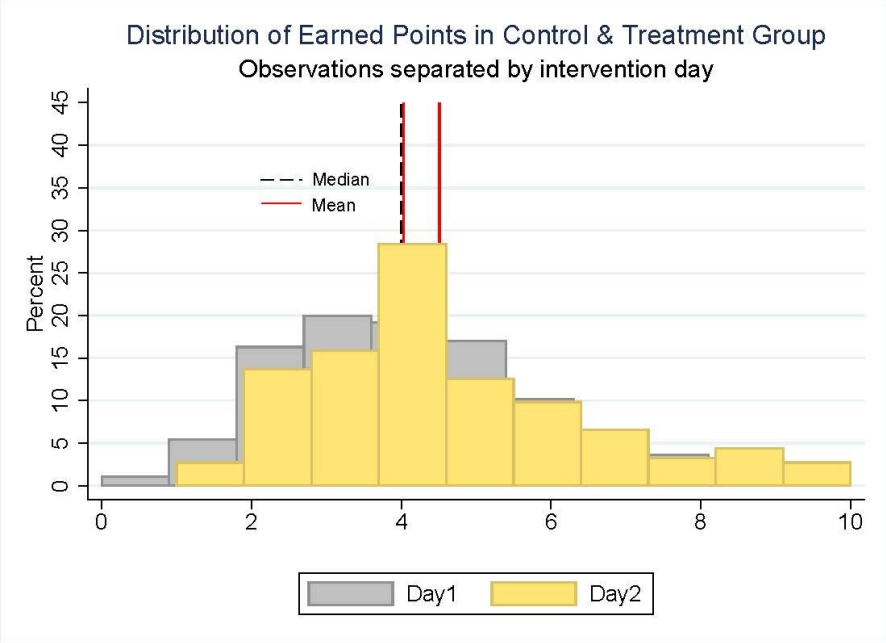


Figure C1: Distribution of earned points (control & treatment group) per intervention day

Table C2: Carryover effects

	Aggregating Evidence Module Score	Impact Evaluation Module Score
Stage 2 (γ)	0.00381 [0.234]	-0.0093 [0.192]
Treatment Group (β)	-0.0909 [0.237]	0.0307 [0.203]
Treat.##Stage 2 (DiD)	-0.112 [0.369]	0.0746 [0.298]
Day 2	-0.108 [0.187]	-0.00298 [0.148]
Female Officer	0.490* [0.192]	0.168 [0.159]
Officer's region of origin (Punjab omitted)	Yes	Yes
Constant	8.306*** [0.225]	7.068*** [0.198]
Observations	452	452
Adjusted R-squared	0.069	0.005

White-Huber robust standard errors in brackets; * p<0.05; ** p<0.01; *** p<0.001

Appendix D – Treatment group analysis

Table D1: Summary statistics treatment group by intervention day

	First intervention day		Second intervention day	
Variable	Mean	Standard Deviation	Mean	Standard Deviation
Individual Average Score (Stage 1)	4.14	1.15	4.3	1.45
Group Score (Stage 2)	4.29	1.8	4.6	2
Super Student Score (Stage 1)	6.55	1.39	6	1.71
Top Student Score (Stage 1)	5.16	1.44	4.96	1.67
Gain (Stage 2 – Stage 1)	0.14	1.49	0.3	1.33
Super Surplus (Stage 1)	2.41	0.96	1.7	0.75
Top Surplus (Stage 1)	1.01	0.8	0.66	0.64
Collaborative Efficiency Score (Stage 2)	0.68	0.31	0.77	0.24
Observations	38		25	

Table D2: Difference in means for treatment group outcomes between both intervention days (t-tests)

Standardized scores: x* = [x-mean(x)]/sd (x)	Mean difference between day 1 and 2	t-statistic
Individual Average Score	-0.122	(-0.47)
Group Score	-0.166	(-0.64)
Super Student Score	0.360	(1.41)
Top Student Score	0.130	(0.50)
Gain (Stage 2 – Stage 1)	-0.109	(-0.42)
Super Surplus	0.748**	(3.10)
Top Surplus	0.467	(1.85)
Collaborative Efficiency	-0.0937	(-1.28)
Observations	63	

* p<0.05; ** p<0.01; *** p<0.001

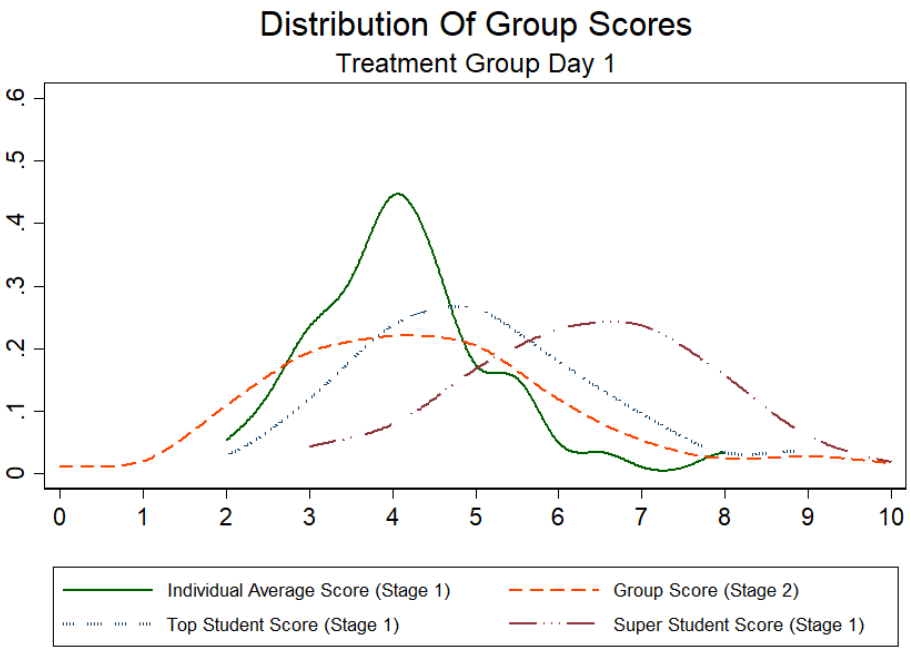


Figure D1: Group scores, first intervention day

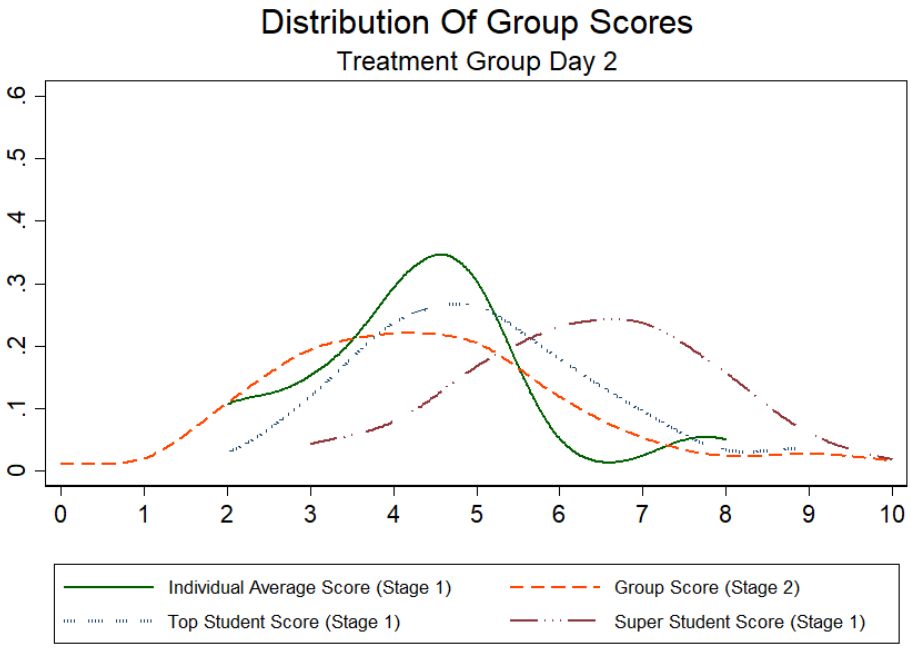


Figure D2: Group scores, second intervention days

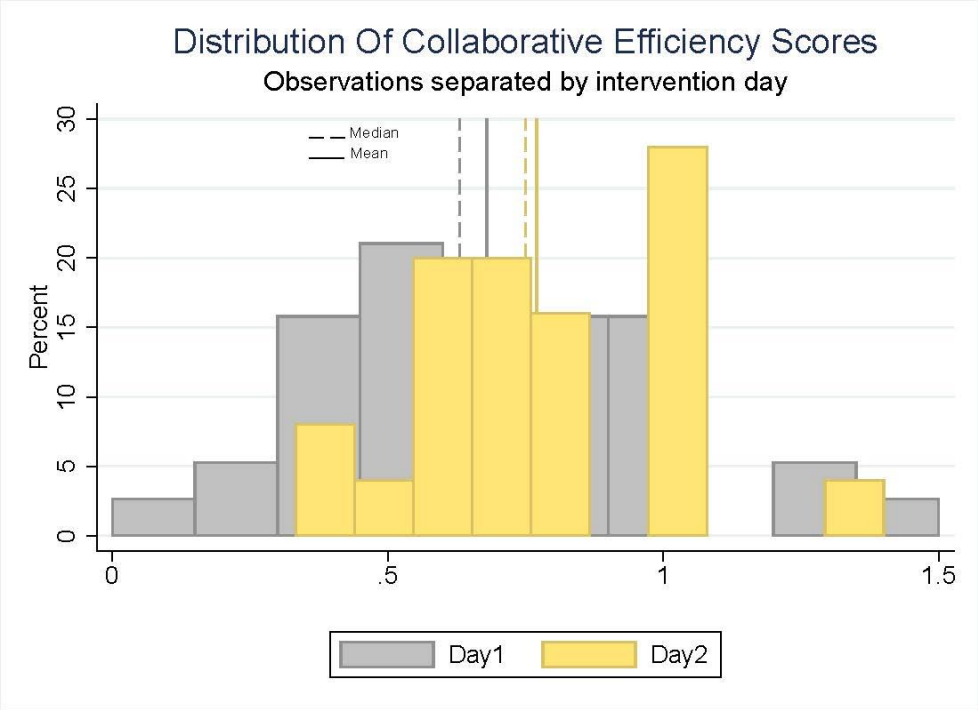


Figure D3: Collaborative efficiency scores, separated by intervention days

Appendix E – Experimental Protocols and Materials

1. Training protocol for CSA and CERP research assistants in Lahore, Pakistan

2.

In-class intervention, November 12 and 13
Civil Services Academy (CSA)

Research Protocol Training

Dr. Laura Metzger
Evidence for Policy Design (EPoD)
Harvard Kennedy School of Government

1

Background Info
What we want to know

Basically, we want to know if people work and learn better in groups or individually. Can we use group work to improve learning outcomes?

To answer this question, we compare the performance of people working in groups with the performance of people working individually.

How do we measure performance? With test scores that will translate into a grade.

2

Intervention Set-Up

START
7:45 AM

END
11:20 AM

Day 1 (n=161), Nov. 12th

Control - Individual Work - n=81	Treatment - Group Work - n=80
Stage 1: individual	Stage 1: individual
Stage 2: individual	Stage 2: groups n=40

3

Intervention Set-Up

START
8:00 AM

END
11:40 AM

Day 2 (n=109), Nov. 13th

Control - Individual Work - n=55	Treatment - Group Work - n=54
Stage 1: individual	Stage 1: individual
Stage 2: individual	Stage 2: groups n=27

4

Schedule – general overview

START
8:00 AM

REGISTRATION DESK - right outside of auditorium

names ordered alphabetically (brackets: A-G/H-N/O-U/V-Z) → officers take envelope with their name (name on post-it, to be removed) → officers take any seat in auditorium

Stage 1: control and treatment together in auditorium

15-minute break: control group remains in auditorium; treatment group moves to two adjacent rooms

Stage 2: control group in auditorium; treatment group in adjacent rooms

END
11:40 AM

5-minute break: regather everyone in auditorium, finish session

5

Rooms and team assignment

Let's assign the auditorium team:

Let's assign the group room 1 team:

Let's assign the group room 2 team:

6

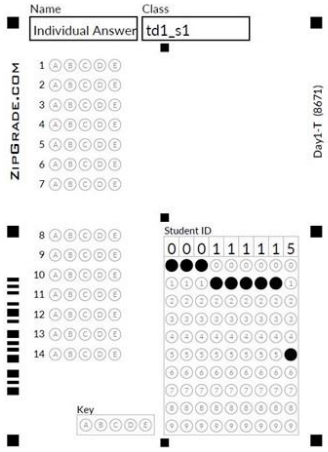
Task material, overview for your information

- Form2_Day1_Task_PRINT2sided.pdf
- Form3_Day1_Treat_PRINT2sided.pdf

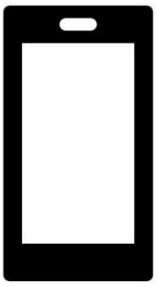
Q1
- a) masdfjasfras
- b) asdfnaoehfr
- c) Asdfkjbad

Qxxx
- a) masdfjasfras
- b) asdfnaoehfr
- c) Asdfkjbas

Q14
- a) masdfjasfras
- b) asdfnaoehfr
- c) Asdfkjbas

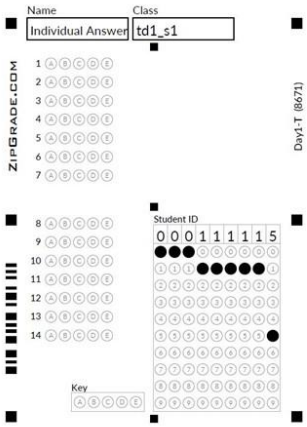


Software & Login Credentials



- We use **ZipGrade** to generate and scan answer sheets for the in-class exercise
- Please install the app on your phone or tablet NOW → go to your AppStore (iphone) or Google play store (android), look for ZIPGRADE
- Login credentials
  Username:
  Password:

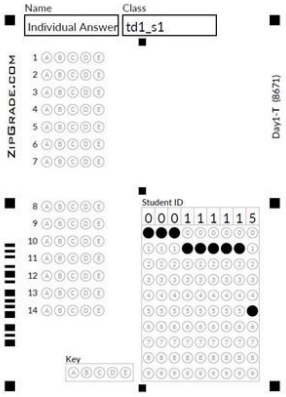
Answer sheet scanning using the right quiz!



Quiz name	Corresponding classes
Day1 Control	cd1_s1, cd1_s2
Day1 Treatment	td1_s1, td1_s2
Day2 Control	cd2_s1, cd2_s2
Day2 Treatment	td2_s1, td2_s2

9

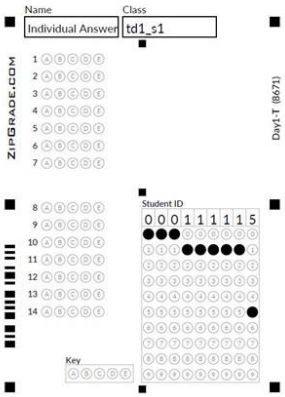
Test scanning exercise



- Let's try a few test scans!
- Circle any bubbles on the sheet
- Hover your phone over the sheet to scan it
- Control group: 2 sheets, Stage 1 (individual sheet) and Stage 2 (individual sheet)
- Treatment group: 2 sheets, Stage 1 (individual sheet) and Stage 2 (group sheet, only one per group!)

10

Technical issues



- sometimes ZipGrade will scan the same sheet twice (happens very fast, often when light is bad) → if it happens, please use the second scan (app will overwrite the first scan)
- if answer aren't circled appropriately, ZipGrade may not recognize the answer and record it as a "blank" → if it happens, please do not try to fill in the bubbles yourself

11

Collecting the material from officers

- Please make sure that – after stage 2 is finished – officers place ALL materials back in the envelopes before you collect it (Handouts, questions, answer sheets) EXCEPT for one specific sheet (see last point), which they are supposed to keep
- We will point out to them that they are not supposed to take any of the material with them
- Handling of "extra-questions-sheet" for officers in the treatment group

12

What if…

…participants try to cheat/ talk to each other?

- remind them calmly and in a friendly tone that talking is not allowed and that another attempt will result in exclusion from the exercise

…participants lose/ completely mess-up their answer sheets?

- we will provide them with a new sheet, but this should only happen in *exceptional* cases

…participants circle more than one bubble for on an answer row?

- ZipGrade will only take the right answer and not penalize an alternative attempt

…participants don't show up or drop out at some point?

- note down ID number displayed on answer sheet and send info to WhatsApp group

…participants don't show up and groups are affected by that?

- note down group number and send info to WhatsApp group if entire group is missing
- note down ID number of present group member and send the info to WhatsApp group;
  - assign present group member to another group if possible, write down that group's number & send info to WhatsApp group!

13

Let's make a WhatsApp group

14

2. Warm up round instructions and questions (identical for control and treatment group and for both intervention days)

Instructions: Warm-Up Round

In the following, we ask you to answer questions based on the information provided in your handout.

This warm-up round will help you to familiarize yourself with the exercise. It is ungraded!

You have **30 minutes to review the information** in the handout **and answer as many questions as you can**. We will then review the answers together.

1. Please read the text carefully.
2. When you are ready, the instructor will start collecting votes on your answers. Simply raise your hand to vote on your preferred answer!

After the warm-up round, we will take a short break and you will receive further instructions to continue working on the exercise.

Please note that **your answers will be recorded anonymously**; it is not possible to identify you based on your answers.

Questions: Warm-Up Round

Question 1: Based on your *previous* knowledge and experience, how effective would you expect the Government of Athana's intervention to be in reducing diarrhea incidence in children under five?

a. Extremely effective
b. Very effective
c. Moderately effective
d. Slightly effective
e. Not effective at all

Question 2: Based on your *previous* knowledge and experience, how effective would you expect the Government of Athana's intervention to be in increasing households' monthly income?

a. Extremely effective
b. Very effective
c. Moderately effective
d. Slightly effective
e. Not effective at all

Question 3: What is the main treatment intervention the Government of Athana plans to undertake?

a. Reduce time needed to fetch water
b. Provide access to a safe water source in target villages
c. Increase individuals' economic productivity
d. None of the above

Question 4: Based on the data ***in the table alone***, each of the three study designs indicates a statistically significant reduction in diarrhea in children under five.

a. True
b. False

Question 5: What are the key criteria the Government used to select the target villages?

a. High incidence of diarrhea in children under five
b. Inadequate access to safe drinking water
c. Relatively high rates of child deaths due to diarrhea
d. All of the above

Question 6: Which of the three study designs have a comparison group (i.e., an approximation of the counterfactual)?

a. Design A
b. Design B
c. Design C
d. All designs have a comparison group

3. Main round instructions for the control group (identical for both intervention days)

Instructions: Main Round

In this round, we ask you to continue answering questions based on the information provided in your handout. Based on the content completed in the online modules and the in-class recap this morning, do your best to answer these questions.

Please do not communicate with others during the entire exercise **including the break**! Communication will result in exclusion from the exercise.

This round has **two stages.**

1. During the **first stage, you will have 30 minutes to work on the questions by yourself**. Please mark your answers on your individual answer sheet. Answer as many questions as you can.

   After the **30 minutes** are over, **we** will **take a 15 minute-break**. Please **remain seated during the break** and **turn your answer sheet over**. An assistant will walk up to your seat and scan your answers.

   After the break is over, you start the second and final stage.

2. During the second stage**, you will have an additional 20 minutes to revise or continue to work on your answers** from the first stage. As in stage 1, you will work by yourself.

   We will provide you with a **fresh sheet for the second stage**. Please **copy your answers from the first stage sheet over** to this fresh sheet.

Grading: Your first stage score will count 0.6 towards your grade while your second stage answers will count 0.4 towards your grade. You can never decrease your stage 1 score, you can only improve it!

- For example, if your stage 1 score is 7 and your second stage score is 6, your final score will be 7
- Another example: if your stage 1 score is 5 and your stage 2 score is 7, your final score will be 5*0.6 + 7*0.4 = 5.8

Once you are finished with the exercise**, place all material back in the envelope and leave it at your seat. We will collect it.** Please **DO NOT take your answer sheets with you,** otherwise we cannot grade the exercise.

During the rest of the in-class session, we will jointly discuss the exercise and the correct answers.

4. Main round instructions for the treatment group (identical for both intervention days)

Instructions: Main Round

In this round, we ask you to continue answering questions based on the information provided in your handout. Based on the content completed in the online modules and the in-class recap this morning, do your best to answer these questions.

This round has **two stages.**

1. During the **first stage**, **you will have 30 minutes to work on the questions by yourself**. Please mark your answers on your individual answer sheet. Answer as many questions as you can.

   Please do not communicate with others during the first stage **including the break**! Communication will result in exclusion from the exercise.

   After the **30 minutes** are over, **we** will **take a 15 minute-break**. During this break, **we will guide you to another room**, where you will complete the second stage.

   Once you arrive in the other room, take a seat at the spot with your group number. Your group number is displayed on your answer sheet.

2. During the second and **final stage, you will have an additional 20 minutes to revise or continue to work on your answers by working in groups of two**.

   Groups are randomly assigned.

   Rules for group work: Please **agree on your answers and mark them on your group answer sheet**. You only need to **hand in one group answer sheet for** the **both of you**.

   Please do not communicate with other *groups* during the exercise! Communication with other groups will result in exclusion from the exercise.

Grading: Your individual stage 1 score will count 0.6 towards your grade while your group stage 2 score will count 0.4 towards your grade. You can never decrease your stage 1 score, you can only improve it!

- For example, if your stage 1 score is 7 and your second stage score is 6, your final score will be 7
- Another example: if your stage 1 score is 5 and your stage 2 score is 7, your final score will be 5*0.6 + 7*0.4 = 5.8

Once you are finished with the exercise**, place all material back in the envelope and leave it at your seat. We will collect it.** Please **DO NOT take your answer sheet with you,** otherwise we cannot grade the exercise.

During the rest of the in-class session, we will jointly discuss the exercise and the correct answers.

Before you leave the auditorium to find your group partner in the designated room, tell us quickly by how much you think group work will improve your grade? Circle the corresponding letter below.

a. 0% to 20%
b. 21% to 40%
c. 41% to 60%
d. 61% to 80%
e. 81% to 100%

5. Task text and questions for the **firs**t intervention day (identical for control and treatment group)

Identifying Reliable Study Results & Expected Program Effects

Please read the following information carefully!

Your colleague's decision problem

The Government of Athana state wants to improve safe drinking water access in districts with a high incidence of diarrhea in children under five and insufficient access to safe drinking water. Diarrhea is a well-known consequence of unsafe drinking water and is harmful to small children. It is estimated that, in targeted districts, about 10% to 12% of deaths in children under five are caused by diarrhea - which is above the national average. In addition to reducing diarrhea in children under five, the government of Athana expects that access to safe drinking water will improve households' income: household members will spend less time fetching water and are less likely to fall ill, both of which can increase their economic productivity.

Government employees gathered evidence from three studies conducted across different states in Norina, the country Athana state is located in. All studies evaluate the impact of improved access to safe drinking water. All three studies evaluate the type of program Athana wants to implement in targeted district villages: installation of hand-pumps, which will provide a safe drinking water source in central locations in villages. The goal of this program will be to reduce diarrhea in children under five and improve household income.

A colleague from Athana **asks you to help her identify the most reliable impact evaluation study results**. This is vital to gain a more precise understanding of expected program effects and to decide whether to move ahead with the intervention.

Your colleague provides you with the evidence she wants to discuss. Please turn the page to review the information.

Three Impact Evaluation Designs

Summary

Program intervention type: Installation of hand pumps serving as a safe water source to beneficiaries.

Target group: Villages in districts with a high diarrhea incidence in children under five and inadequate access to safe drinking water.

Outcome indicators:

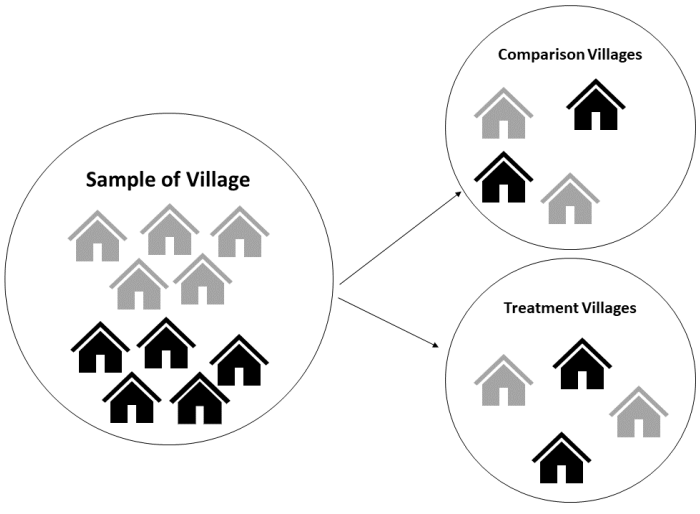
- Incidence of diarrhea in children under five, two months after installing the hand pumps.
- Household income two months after installing the hand pumps.

Assumptions you can make:

- All hand pumps were installed and are functional.
- Drinking water from the hand pumps is safe.
- Any villager can go to a hand pump, obtain water, and return to their home within 30 minutes.
- The gathered studies are relevant for your colleague's decision problem, since they were conducted in similar contexts across the same country.

Design A: Randomized Evaluation

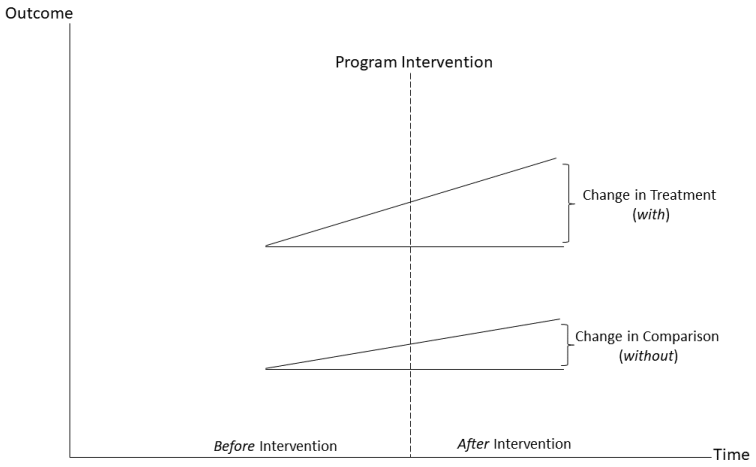
This study randomly assigns villages to treatment and comparison groups. This ensures that there are no differences, on average, between these two groups in terms of their characteristics. The only difference is whether they receive a hand-pump or not. In other words, the study measures the *difference in outcomes* between treatment and comparison villages that are *randomly assigned* to the program.



Impact as determined by Design A: Any *difference* in outcomes between *randomly assigned* treatment and comparison villages will be attributed to the program.

Design B: Difference-in-differences

This study combines before-after and with-without comparisons. It compares the change in outcomes over time between the treatment villages and comparison villages in a neighboring district. In other words, the study measures whether the outcome changes by more or less in the treatment villages compared to the comparison villages.

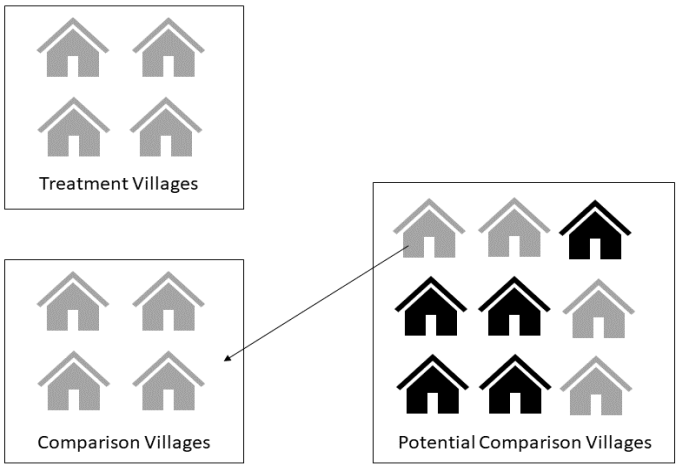


Impact as determined by Design B: Any *difference in the change* in outcomes between treatment and comparison group over time will be attributed to the program. Or, expressed differently: the change in treatment group outcomes *minus* the change in control group outcomes *equals* the impact attributed to the program.

Note: This graph only illustrates the study design in theory, it does not represent what actually happened to the diarrheal outcomes in the study described in the text.

Design C: Matching

This study matches and compares treatment and comparison group villages based on their similarity in *observed socio-demographic* characteristics. In other words, the study manually constructs a comparison group from villages that are similar to treatment villages along a set of *predetermined* variables and then compares the *differenc*e in *outcomes between* the groups. Examples of variables used to match include (but are not limited to) average income, education level of household head, household size, and employment rate. Important for this design's validity is that the variables used for matching are not affected by the treatment. Therefore, one should only use variables whose values were observed before the treatment was administered or whose values remain constant over time.



Impact as determined by Design C: Any *difference* in outcomes between *matched* treatment and comparison villages will be attributed to the program.

Average Effects:

The table below shows the average effect of installing hand pumps on each outcome of interest in treatment group villages estimated by each study. Except for the estimated effect of hand-pumps on household income shown by study C, all estimated effects are statistically significant.

	Diarrhea in children under five years, past two weeks	**Income per household, monthly**
Design A: Randomized Evaluation	decreased by about 21%	increased by about 4%
Design B: Difference-In-Difference	decreased by about 33%	increased by about 6%
Design C: Matching	decreased by about 36%	increased by about 3% (not statistically significant)

Identifying Reliable Study Results & Expected Program Effects

Questions: Main Round

Question 1: Consider Study Design A in the handout. Which alternative explanation is not ruled out by this evaluation?

a. Systematic differences in mean income, household size, and education between treatment and control villages at project start
b. A region-wide hand washing campaign that affects treatment and control villages alike
c. Differences in out-migration rates between treatment and control villages over the course of the study

Question 2: Consider Study Design A again. The study report states that villages in the program region are very heterogeneous. In your opinion, does this pose a threat to the study design?

a. Yes, it does pose a threat to the study design.
b. No, it does not pose a threat to the study design.

Question 3: Consider Study Design B. Which alternative explanation is ruled out by this evaluation?

a. Changes in child health policies in one district but not the other
b. State-level income subsidies for poor families introduced during the intervention
c. A large sanitation project started by the World Bank in one district but not the other

Question 4: Consider Study Design B again. The study shows time trends on various health outcomes in comparison and treatment villages during the three years prior to project start. Comparison and treatment villages seem to follow parallel trends for diarrhea incidence in children under five but seem to follow different trends for most other health outcomes. How does this information affect your assessment of the comparison group's quality?

a. I would rate the comparison group's quality lower.
b. I would rate the comparison group's quality higher.
c. I would rate the comparison group's quality the same.

Question 5: Consider Study Design C. Which alternative explanation is not ruled out by this evaluation?

a. Systematic differences in political ties between treatment group villages and the state's ruling party
b. Systematic differences in village size between comparison and treatment villages
c. Systematic difference in household assets between comparison and treatment villages (bicycle, radio, cell phones)

Question 6: Consider Study Design C again. Below is a subset of variables about individual characteristics that were available for the study. All variables were measured *after* the study (e.g., example "household income" refers to household income *after* the intervention). Which one would be most appropriate to be used in constructing the comparison group?

a. A child's gender
b. Employment status (0=unemployed, 1=employed)
c. Household income

Question 7: Imagine that, in the case of Study Design A, the government started a handwashing campaign in control and treatment villages a few months into the project. In this case, what can you say about the reported effect of hand pumps on diarrhea in children under five?

a. The study likely overestimates the effect
b. The study likely underestimates the effect
c. The study likely correctly estimates the effect

Question 8: Imagine that, in the case of Study Design B, comparison villages experienced a rainfall shortage that worsened drinking water access and led to an increase in waterborne diseases including diarrhea. In this case, what can you say about the reported effect of hand pumps on diarrhea in children under five?

a. The study likely overestimates the effect
b. The study likely underestimates the effect
c. The study likely correctly estimates the effect

Question 9: Imagine that, in the case of Study Design C, treatment villages were exposed to a cholera outbreak in the previous year which had sensitized villagers towards the dangers of falling ill from diarrhea. In this case, what can you say about the reported effect of hand pumps on diarrhea in children under five?

a. The study likely overestimates the effect
b. The study likely underestimates the effect
c. The study likely correctly estimates the effect

To answer questions 10 through 14, please consider your answers to questions *7, 8*, and *9 only* and assume that each scenario in those questions actually occurred.

Question 10: What study design, do you think, has the most reliable comparison group?

a. Study Design A
b. Study Design B
c. Study Design C

Question 11: After having assessed the evidence, how effective would you expect the described intervention to be in reducing diarrhea incidence in children under five?

a. Extremely effective
b. Very effective
c. Moderately effective
d. Slightly effective
e. Not effective at all

Question 12: After having assessed the evidence, how effective would you expect the described intervention to be in increasing households' monthly income?

a. Extremely effective
b. Very effective
c. Moderately effective
d. Slightly effective
e. Not effective at all

Question 13: Do you recommend your colleague to move ahead with the intervention?

a. Yes
b. No

Question 14: What is the level of confidence you have in your answer to question 10?

a. High
b. Medium
c. Low

6. Task text and questions for the **second** intervention day (identical for control and treatment group)

Identifying Reliable Study Results & Expected Program Effects

Please read the following information carefully!

Your colleague's decision problem

The Government of Athana state wants to improve safe drinking water access in districts with a high incidence of diarrhea in children under five and insufficient access to safe drinking water. Diarrhea is a well-known consequence of unsafe drinking water and is harmful to small children. It is estimated that, in targeted districts, about 10% to 12% of deaths in children under five are caused by diarrhea - which is above the national average. In addition to reducing diarrhea in children under five, the government of Athana expects that access to safe drinking water will improve households' income: household members will spend less time fetching water and are less likely to fall ill, both of which can increase their economic productivity.

Government employees gathered evidence from three studies conducted across different states in Norina, the country Athana state is located in. All studies evaluate the impact of improved access to safe drinking water. All three studies evaluate the type of program Athana wants to implement in targeted district villages: installation of hand-pumps, which will provide a safe drinking water source in central locations in villages. The goal of this program will be to reduce diarrhea in children under five and improve household income.

A colleague from Athana **asks you to help her identify the most reliable study results**. This is vital to gain a more precise understanding of expected program effects and to decide whether to move ahead with the intervention.

Your colleague provides you with the evidence she wants to discuss. Please turn the page to review the information.

Three Impact Evaluation Designs

Summary

Program intervention type: Installation of hand pumps serving as a safe water source to beneficiaries.

Target group: Villages in districts with a high diarrhea incidence in children under five and inadequate access to safe drinking water.

Outcome indicators:

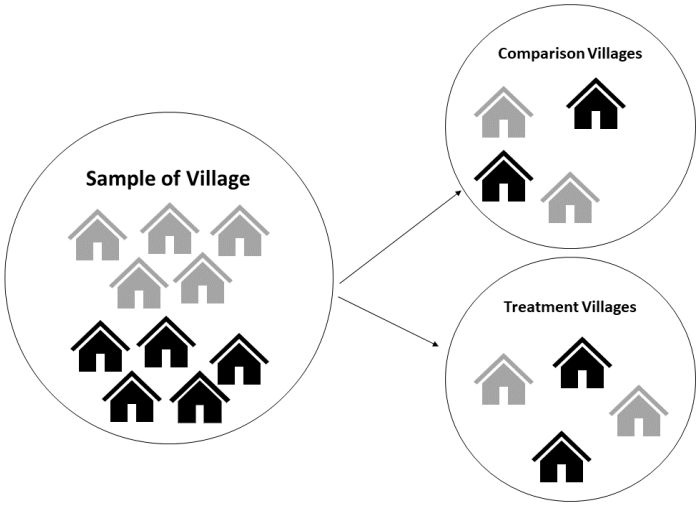
- Incidence of diarrhea in children under five two months after installing the hand pumps.
- Household income two months after installing the hand pumps.

Assumptions you can make:

- All hand pumps were installed and are functional.
- Drinking water from the hand pumps is safe.
- Any villager can go to a hand pump, obtain water, and return to their home within 30 minutes.
- The gathered studies are relevant for your colleague's decision problem, since they were conducted in similar contexts across the same country.

Design A: Randomized Evaluation

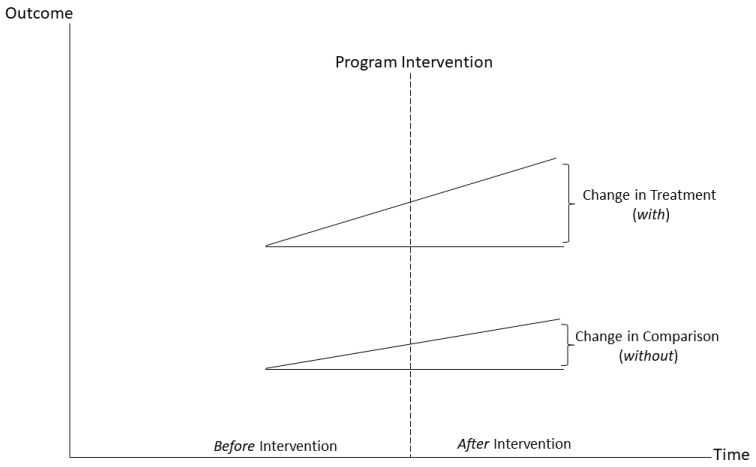
This study randomly assigns villages to treatment and comparison groups. This ensures that there are no differences, on average, between these two groups in terms of their characteristics. The only difference is whether they receive a hand-pump or not. In other words, the study measures the *difference in outcomes* between treatment and comparison villages that are *randomly assigned* to the program.



Impact as determined by Design A: Any *difference* in outcomes between *randomly assigned* treatment and comparison villages will be attributed to the program.

Design B: Difference-in-differences

This study combines before-after and with-without comparisons. It compares the change in outcomes over time between the treatment villages and comparison villages in a neighboring district. In other words, the study measures whether the outcome changes by more or less in the treatment villages compared to the comparison villages.

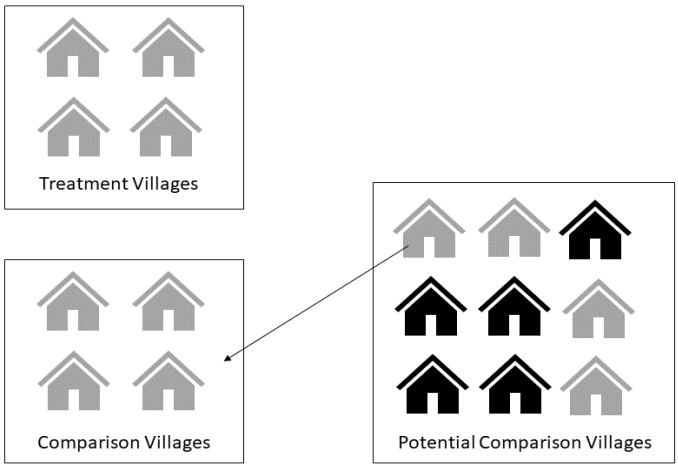


Impact as determined by Design B: Any *difference in the change* in outcomes between treatment and comparison group over time will be attributed to the program. Or, expressed differently: the change in treatment group outcomes *minus* the change in control group outcomes *equals* the impact attributed to the program.

Note: This graph only illustrates the study design in theory, it does not represent what actually happened to the diarrheal outcomes in the study described in the text.

Design C: Matching

This study matches and compares treatment and comparison group villages based on their similarity in *observed socio-demographic* characteristics. In other words, the study manually constructs a comparison group from villages that are similar to treatment villages along a set of predetermined variables and then compares the *differenc*e in *outcomes between* the groups. Examples of variables used to match include (but are not limited to) average income, education level of household head, household size, and employment rate. Important for this design's validity is that the variables used for matching are not affected by the treatment. Therefore, one should only use variables whose values were observed before the treatment was administered or whose values remain constant over time.



Impact as determined by Design C: Any *difference* in outcomes between *matched* treatment and comparison villages will be attributed to the program.

Average Effects:

The table below shows the average effect of installing hand pumps on each outcome of interest in treatment group villages estimated by each study. Except for the estimated effect of hand-pumps on household income shown by study A, all estimated effects are statistically significant.

	Diarrhea in children under five years, past two weeks	**Income per household, monthly**
Design A: Randomized Evaluation	decreased by about 19%	increased by about 2% (not statistically significant)
Design B: Difference-In-Difference	decreased by about 24%	increased by about 4.5%
Design C: Matching	decreased by about 29.5%	increased by about 5%

Identifying Reliable Study Results & Expected Program Effects

Questions: Main Round

Question 1: Consider Study Design A. Which alternative explanation is not ruled out by this evaluation?

a. Systematic differences in mean income, household size, and education between treatment and control villages at project start
b. A region-wide hand washing campaign affecting control villages
c. Similar out-migration rates in treatment and control villages over the course of the study

Question 2: Consider Study Design A again. The study report states that villages in the program region are very homogeneous. In your opinion, does this pose a threat to the study design?

a. Yes, it does pose a threat to the study design.
b. No, it does not pose a threat to the study design.

Question 3: Consider Study Design B. Which alternative explanation is ruled out by this evaluation?

a. A state-wide change in child health policies
b. Income subsidies for poor families introduced in one district but not the other
c. A large sanitation project started by the World Bank covering treatment districts

Question 4: Consider Study Design B again. The study shows time trends on various health outcomes in comparison and treatment villages during the three years prior to project start. Comparison and treatment villages seem to follow similar trends for diarrhea incidence in children under five, and seem to follow similar trends for most other health outcomes. How does this information affect your assessment of the comparison group's quality?

a. I would rate the comparison group's quality lower.
b. I would rate the comparison group's quality higher.
c. I would rate the comparison group's quality the same.

Question 5: Consider Study Design C. Which alternative explanation is not ruled out by this evaluation?

a. Systematic differences in political ties between treatment group villages and the state's ruling party
b. Systematic differences in village size between comparison and treatment villages
c. Systematic difference in household assets between comparison and treatment villages (bicycle, radio, cell phones)

Question 6: Consider Study Design C again. Below is a list of variables the researchers had available to measure diarrhea in children under five. Which one would be most reliable to measure diarrhea?

a. Number of incidents of diarrhea during the first 1 week after project start (surveyed once; reported by the mother 1 week after the project start)
b. Number of incidents of diarrhea during the first 5 months after project start (surveyed once; reported by the mother 5 months after the project start)
c. Number of incidents of diarrhea during the first and second month after project start (surveyed twice, once in the first month and once in the second month; reported by the mother)

Question 7: Imagine that, in the case of Study Design A, the government started a handwashing campaign in control villages a few months into the project. In this case, what can you say about the reported effect of hand pumps on diarrhea in children under five?

a. The study likely overestimates the effect
b. The study likely underestimates the effect
c. The study likely correctly estimates the effect

Question 8: Imagine that, in the case of Study Design B, comparison and treatment villages experienced a rainfall shortage that worsened drinking water access and led to an increase in waterborne diseases including diarrhea. In this case, what can you say about the reported effect of hand pumps on diarrhea in children under five?

a. The study likely overestimates the effect
b. The study likely underestimates the effect
c. The study likely correctly estimates the effect

Question 9: Imagine that, in the case of Study Design C, treatment villages were exposed to a cholera outbreak in the previous year which had sensitized villagers towards the dangers of falling ill from diarrhea. In this case, what can you say about the reported effect of hand pumps on diarrhea in children under five?

a. The study likely overestimates the effect
b. The study likely underestimates the effect
c. The study likely correctly estimates the effect

To answer questions 10 through 14, please consider your answers to questions *7, 8*, and *9 only* and assume that each scenario in those questions actually occurred.

Question 10: What study design, do you think, has the most reliable comparison group?

a. Study Design A
b. Study Design B
c. Study Design C

Question 11: After having assessed the evidence, how effective would you expect the described intervention to be in reducing diarrhea incidence in children under five?

a. Extremely effective
b. Very effective
c. Moderately effective
d. Slightly effective
e. Not effective at all

Question 12: After having assessed the evidence, how effective would you expect the described intervention to be in increasing households' monthly income?

a. Extremely effective
b. Very effective
c. Moderately effective
d. Slightly effective
e. Not effective at all

Question 13: Do you recommend your colleague to move ahead with the intervention?

a. Yes
b. No

Question 14: What is the level of confidence you have in your answer to question 10?

a. High
b. Medium
c. Low