
DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY


# Checking all the boxes: Encouraging the adoption of behavioral science at an international organization through a recruitment pilot using asynchronous video interviews

## Citation

Nakagawa, Jun. 2021. Checking all the boxes: Encouraging the adoption of behavioral science at an international organization through a recruitment pilot using asynchronous video interviews. Doctoral dissertation, Harvard T.H. Chan School of Public Health.

## Link

https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37370666

## Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material (LAA), as set forth at

https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFlOWI2M2EwYTg

## Accessibility

https://accessibility.huit.harvard.edu/digital-accessibility-policy

## Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story

This Doctoral Thesis, *Checking all the boxes: Encouraging the adoption of behavioral science at an international organization through a recruitment pilot using asynchronous video interviews*, presented by Jun Nakagawa and Submitted to the Faculty of The Harvard T.H. Chan School of Public Health in Partial Fulfilment of the Requirements for the Degree of Doctor of Public Health, has been read and approved by:

K. Viswanath

Dr. Kasisomayajula Viswanath

Linda A Cyr

Dr. Linda Cyr

C. Kreatsoulas

Dr. Catherine Kreatsoulas

October 12, 2021

# Checking all the boxes: Encouraging the adoption of behavioral science at an international organization through a recruitment pilot using asynchronous video interviews

Jun Nakagawa

A Doctoral Thesis Submitted to the Faculty of

The Harvard T.H. Chan School of Public Health

in Partial Fulfillment of the Requirements

for the Degree of *Doctor of Public Health,*

Harvard University

Boston, Massachusetts.

March 2022

Thesis Advisor: Dr. Kasisomayajula Viswanath Jun Nakagawa

Checking all the boxes: Encouraging the adoption of behavioral science at an international organization through a recruitment pilot using asynchronous video interviews

## Abstract

Traditional recruitment practices are prone to unconscious biases that can disadvantage minority groups. To combat this, many organizations have publicly committed to more inclusive workplaces by changing their recruitment policies or investing in digital employment selection tools. Ensuring fairer recruitment practices has important downstream public health implications as unemployment and discriminatory work environments can contribute to poor health outcomes such as adverse effects on one's mental and physical health. However, adopting new policies and technology is not a silver bullet because without proper design and evaluation within the host organization, there is a potential danger of perpetuating existing discriminatory norms. In response to these challenges, the objective of the Doctoral Project was to conduct a pilot with HR professionals to evaluate whether using different mediums of asynchronous video interviews (AVIs) affected the selection of entry-level positions. To carry out this objective, I leveraged principles from Mark Moore's Strategic Triangle where I focused on gaining legitimacy and support, developing operational capacity, and creating public value in the host organization. In this dissertation, I outline key findings from the analysis and share lessons learned from being on the host organization's first behavioral design team.

Table of Contents

*Abstract* ..... *ii*

*List of Figures with Captions* ..... *4*

*List of Tables* ..... *vi*

*Acknowledgments* ..... *vii*

*Executive Summary* ..... *8*

*Chapter 1: Introduction* ..... *11*

*Chapter 2: Analytical Platform* ..... *14*

**Public health implications of fair recruitment** ..... **14**

**Behavioral biases in recruitment practices** ..... **16**

**Digital recruitment tools and biases** ..... **17**

*Chapter 3: Theory of Change* ..... *19*

**Political Work** ..... **20**

**Managerial Work** ..... **24**

**Imaginative Work** ..... **26**

**Conclusion** ..... **27**

*Chapter 4: Analysis* ..... *29*

**Introduction** ..... **29**

Background: Recruiting young talent at the Organization ..... 29

Rationale ..... 30

Objectives ..... 31

Hypotheses ..... 32

Methods ..... 36

Statistical Analysis ..... 40

**Results** ..... **41**

Participants ..... 41

Descriptive data ..... 42

Main results ..... 43

Additional analysis ..... 48

**Discussion** ..... **49**

Key results ..... 49

Limitations ..... 51

**Implications for public health research and practice** ..... **51**

**Future directions** ..... **53**

*References* ..... *60*

*Appendices* ..... *69*

*Appendix 1: Summary of gift reporting campaign* ..... **69**

*Appendix 2: Table showing the number of included candidates for each job description* ..... **75**

*Appendix 3: Interview questions and grading rubric* ..... **76**

## List of Figures with Captions

**Figure 1:** Mark Moore's Strategic Triangle

**Figure 2:** Organigramme of the Executive Department (ED)

**Figure 3:** Phases to increase informal authority (Adapted form Heifetz, 2009)

**Figure 4:** Stakeholder map

**Figure 5:** Summary of how the Project was secured using the Moore's Strategic Triangle framework

**Figure 6:** Five stages of the 2010-18 YPP and YAP recruitment process

**Figure 7:** Five stages of the 2018-19 YPP and YAP recruitment process

**Figure 8:** Flow chart showing inclusion criteria of data sample

## List of Tables

**Table 1:** Inherited study design (pilot)

**Table 2:** Demographic characteristics of all included candidates and raters of the hiring committee

**Table 3:** Interview characteristics by medium and shared demographics

**Table 4:** Linear Regression outputs for Models 1-4

**Table 5:** Legend that describes Models 1-4

**Table 6:** Linear regression output to assess if the association between demographics (candidate's gender, candidate's nationality, rater's gender, and rater's nationality) and mean interview score differed by medium subgroup (transcript, audio, and video)

## Acknowledgments

It has been a privilege to work with my committee members: Drs. K. Vish Viswanath, Linda Cyr, and Catherine Kreatsoulas. Dr. Viswanath's wealth of knowledge and passion for health disparities were instrumental in identifying the possible public health implications related to the Doctoral Project (referred to as the Project). I also appreciate Dr. Cyr for always challenging me to think about the "so what?" so that I do not lose sight of the practical applications and real-life learnings. Finally, I am grateful to Dr. Kreatsoulas for her continuous encouragement and patience. She has taught me more about statistics in the past year than all my coursework put together. I would also like to extend my gratitude to the DrPH program, especially to Rick Siegrist for his support throughout my journey to find a Project, Gary Williams for his help in navigating the administrative process, and members of the third DrPH cohort who encouraged me to keep going even when it seemed impossible to find a dissertation topic.

The Project would not have been possible without the host organization. I am grateful to senior leadership of the Executive Department (ED), Human Resource (HR) staff, interns, and the Head of Strategy for taking a chance on me. If not for their willingness to take a risk and reflect on the status quo, the Project would never have been possible. I also appreciate my colleagues on the original behavioral design team who were instrumental in advancing the work and delivering on the Project.

I have also been lucky to have been surrounded by supportive friends and family, most notably my parents, Hiroshi and Hideko Nakagawa, who have encouraged me throughout my journey. I would also like to thank actor Andy Buckley, aka "David Wallace," whose motivational speech brought a smile to my face just moments before my dissertation defense. Finally, I would like to express my gratitude to my husband, Jamie Macfarlane, whose continuous love, support and sense of humor helped me through every stage of the process.

## Executive Summary

For decades, the business world has long recognized the "business case" of a diverse and inclusive workforce such as profits and productivity, and as a result, a growing number of corporations have worked to address the challenges of unconscious bias in the workplace (Azoulay et al., 2011; Reagans & Zuckerman, 2001; Robinson & Dechant, 1997). Unconscious bias is defined as an "unconscious belief about a particular category of people…[that] allows for quick, sometimes inaccurate, processing of information" (Poppenhaeger, 2017). Although there are benefits of relying on short-cuts to make decisions, they can also adversely affect important ones such as hiring decisions. For example, a hiring manager may unconsciously rely on stereotypes that "amplify systematic differences between groups" even if, in reality, the differences are insignificant to determine the capacity of an applicant to do the job (Bordalo et al., 2016). At scale, assuming that the decision-makers themselves are homogenous, this type of decision-making can contribute to a homogenous workforce.

In public health, the discourse on the "public health case" of a diverse and inclusive workforce has only recently gained traction (Alsan et al., 2019). In the recent years, more emphasis has been placed on a employing a diverse workforce that is representative of the communities they serve and support. Organizations that employ a diverse workforce are in a better position to serve communities where they are needed, create systems to support those needs and supply a greater variety of effective solutions to ultimately address health disparities (Coronado et al., 2020). The benefits of addressing unconscious bias for the public health field uniquely differ from the business realm in that a more diverse public health workforce leads to better access to and quality of care for disadvantaged populations (Meghani et al., 2012); a greater diversity of experiences can lead to more innovative public health approaches (Kolev et al., 2019); and better evidence and training related to healthy equity facilitates improvement in public health outcomes (AAMC, 2021). Reaping the benefits of a more diverse public health

workforce can have long-term positive effects not only for the organization but also for the community that its serves.

One way to achieve a more diverse public health workforce is addressing unconscious bias when recruiting new employees in an organization. Although there have been several solutions that have gained popularity such as removing demographic characteristics of applicants without testing solutions within the context of the organization it is difficult to know whether they are preventing or contributing to a diverse workforce (Bohnet, 2016). The objective of my DrPH dissertation was to work with Human Resources (HR) of the host organization (referred to as the "Organization") to identify an opportunity to evaluate one of their existing recruitment tools and understand if there is potential bias that favors one demographic group based on non-merit considerations. To carry out this work, I primarily used Mark Moore's Strategic Triangle as my theory of change framework (Cels et al., 2012). By leveraging this leadership framework, I was able to secure permission to implement the Project through:

1. Secure buy-in from HR and senior stakeholders
2. Manage stakeholder expectations
3. Articulate the vision that was tailored to each stakeholder

At the end of this process, I was able to secure a DrPH Project (referred to as the "Project") where I conducted preliminary analysis on whether demographics characteristics, specifically the gender and nationality of the candidate and HR professional ("rater"), influenced the evaluation score of the candidate. In addition, I examined whether interviews where the rater and candidate shared the same gender or nationality influenced the mean interview score. Finally, I looked at the effect of mode of the interview ("medium") on the evaluation score. The analysis showed that overall, there was a bias in favor of interviews with male candidates and interviews with candidates from highly represented countries. However,

shared demographics between the rater and candidate did not influence the mean interview score and these findings did not differ across mediums. The results suggested that HR staff did not have a preference towards others who share similar background and did not make unfavorable hiring decisions towards people who are dissimilar to them. Finally, the analysis showed that richer mediums (video and audio interviews) where the rater has access to more social cues were more favored than transcript interviews. Although the results of the Project did not directly change the recruitment process of the Organization, it did raise the discussion among senior stakeholders to of how the HR can more regularly measure current recruitment metrics to ultimately improve the Organization's decision-making.

## Chapter 1: Introduction

A diverse workforce is an ideal that many organizations across corporations, government, and academic institutions aspire to (Robinson & Dechant, 1997; Bilimoria et al., 2008; Østergaard et al., 2011). Diversity in the workplace refers to an organization that intentionally employs a workforce comprised of individuals with a range of characteristics, such as gender, religion, race, age, ethnicity, sexual orientation, education, and other attributes that has a varied representation of a population (Saxena, 2014). The business case for diversity has been widely accepted as countless studies have shown that diversity leads to better business outcomes (Azoulay et al., 2011), improved levels of productivity (Reagans & Zuckerman, 2001), and overall higher returns on investments (Hunt et al., 2015). Diverse teams have also been shown to produce more influential research and are more innovative because of the greater number of perspectives (Tasheva & Hillman, 2018). In theory, when teams are set up for diversity, they can bring different ideas to the table, avoid groupthink, and can solve more complex problems. To put this into practice, an increasing number of organizations have prioritized developing more diverse organizations by emphasizing objective measures of candidates' ability and past performance (Reuben et al., 2014), building institutional support for equality (Monroe et al., 2008), increasing the diversity of evaluators (Kunze & Miller, 2017), adopting accountability and transparency procedures (Castilla, 2015), and hiring Chief Diversity Officers to lead their inclusive initiatives (Hewlett, 2007).

Yet, progress is slow. Despite the widely accepted value of diversity and, in parallel, the awareness of discrimination over recent decades, meta-analyses have shown no evidence of a decline in discrimination against black applicants in the labor market since the late 1980s and only slight declines for Latino job seekers (Quillian et al., 2017). Similar effects persist across genders, with one study indicating that even top-tier academic faculty inadvertently rate female candidates for STEM positions lower than identically qualified male candidates (Moss-

Racusin et al., 2012). Several studies have highlighted a contributor to this continued disparity is unconscious bias:

> A positive or negative unconscious belief about a particular category of people. This allows quick, but sometimes inaccurate, processing of information...Biases are not the same as discrimination; discrimination can happen if a person actually acts on their biases...If biases go unchecked, they can have multiple detrimental effects for groups against which negative biases exist, for example in performance evaluations, hiring, and career progression (Poppenhaeger, 2017).

Traditional recruitment and hiring practices are prone to unconscious bias. For example, Human Resource (HR) professionals still continuously rely on requiring applicants to submit resumes. Research shows the average recruiter takes fewer than ten (10) seconds to scan a resume and decide whether to call someone back for an interview (Sanburn, 2012). Although these quick judgments save recruiters time and money, six seconds is not long enough for the average brain to fully digest and analyze an entire resume. Resume scanning enables an environment where people rely on automated or unconscious thinking, "System 1" (Kahneman, 2003), and implicit cues (i.e., white vs ethnic-sounding names) to drive hiring decisions (Derous & Ryan, 2019). While fair recruitment and hiring practices are not the only drivers of reducing gender and racial inequality, they contribute to a system that makes it more difficult for marginalized groups to achieve positive outcomes in their professional and personal lives.

To determine what "fair" recruitment and hiring practice looks like, I looked to two legal definitions used in various countries, including the United States, to describe workplace discrimination: disparate treatment and disparate impact. Disparate treatment occurs when a candidate is affected by an intentional act of overt discrimination, while disparate impact occurs when a hiring practice disproportionately excludes candidates based on protected demographic characteristics (Barocas, Solon; Selbst, Andrew D, 2016; Roberts, 2010). Using both definitions, for this project's purposes, a fair hiring practice is when a candidate is not

directly and intentionally singled out for discriminatory treatment and when the selection process does not disproportionately disadvantage or favor members of any demographic group.

For the Doctoral Project (referred to as the Project), the primary goal was to identify and evaluate potential areas of unconscious bias in the recruitment process of the host organization (referred to as the Organization). A secondary goal was to use the findings of the Project to influence the hiring practice of the Organization and ultimately encourage a more representative public health workforce. The Organization is a multilateral organization that influences health policies of its member countries. Shedding light on improving diversity and inclusivity in influential organizations like the Organization could ultimately impact downstream public health outcomes. In the following chapters, I will outline how I found opportunities of evaluation and what I learned from the research and my own experiences. Chapter 2 (Analytical Platform) looks at the public health implications of fair recruitment practices and outlines best practices in the behavioral sciences to address biases in hiring practices. Chapter 3 describes how I used Mark Moore's Strategic Triangle (Moore, 2000) as a theory of change framework to get buy-in from the Organization to evaluate their recruitment process. Chapter 4 provided background of the pilot, inherited study design, hypotheses, results, and discussion.

## Chapter 2: Analytical Platform

### Public health implications of fair recruitment

Fair recruitment that results in a more diverse workforce has significant public health implication. First, employers of the public health workforce can receive the usual benefits of diversity (e.g., productivity, innovation) as previously mentioned as well as directly influence health outcomes. The public health workforce is composed of those who work for "official public health agencies at all levels of government, community-based, and voluntary organizations with a health promotion focus, the public health-related staff of hospitals and healthcare systems, and a range of others in private industry, government, and the voluntary sector" (Tilson & Gebbie, 2004). A public health workforce representing the society that it serves can contribute to positive health outcomes, especially for minority groups and underserved populations. An increasing number of studies have found that public health agencies that employ a diverse workforce are much better positioned to implement more tailored approaches in underprivileged communities where they are the most needed, create systems to support the communities' needs, and supply more effective and culturally competent solutions to ultimately address health disparities (Bouye et al., 2018; Cohen et al., 2002; Coronado et al., 2020). For example, a local public health department with a diverse workforce that represents the community that it serves is more likely to employ localized strategies such as translating public health information to languages other than English (Coronado et al., 2020). On the other end of the spectrum, public health organizations that employ a homogenous workforce can have serious implications such as low quality healthcare services to minority groups and ultimately contribute to health disparities. For example, field studies have shown that hospitals that a homogenous workforce may be less likely to prescribe HIV treatment to African-Americans (Bogard et al., 2000) or refer cardiac catheterization to women (Kreatsoulas et al., 2010). There is compelling evidence that one way to provide more

evidence-based care for minority patients is by having more minority physicians (Smedley et al., 2001; The Spotlight Team, 2017). Furthermore, countless studies have consistently shown that minority physicians are more likely to choose primary care specialties and practice among the poor in underserved areas (Thurmond & Kirch, 1999; *Missing Persons: Minorities in the Health Professions*, 2004; Bouye et al., 2018b). Thus, in the most direct sense, a more representative public health workforce can potentially improve the health outcomes of minority groups or underserved communities.

Fairer hiring practices can also impact the design and delivery of more inclusive policies. A critical component of translating inclusive growth is having diverse representation in public bodies that influence the design of more inclusive policies. This requires the capacity of public bodies to employ a more diverse and representative workforce that can assess broader viewpoints, needs, preferences, and concerns of a diverse range of stakeholders including underserved populations instead of capturing those from a specific demographic group. If more inclusive recruitment can lead to a more representative public health workforce, then this diversity can be reflected in the decision-making process of the Organization and its member governments, ultimately affecting public health outcomes.

Finally, for typical workplaces, not just public health organizations, employment processes and conditions directly impact the health and well-being of their employees and communities. Traditionally, workplace health has been concerned with ensuring that workers are safe from exposure to physical or chemical hazards. In recent decades, workplace health has expanded to psychosocial workplace development to include health inequalities through stress-related processes that stem from structural or institutional discrimination (McCluney et al., 2018). Employment has also been recognized by the World Health Organization's commission on social determinants of health as a major route whereby individuals gain status and resources that are beneficial to health and well-being (Marmot & Bell, 2010). A number

of studies have found that job insecurity and unemployment not only limit income and access to material resources for health but also are contributors to poor health outcomes, including adverse effects on mental and physical health (Caroli & Godard, 2016; Ferrie, 2002; Janlert & Hammarström, 2009). Thus, ensuring organizations do not intentionally or unintentionally engage in discriminatory hiring practices can positively influence health outcomes for the organizations and their communities.

## Behavioral biases in recruitment practices

How can organizations ensure fair recruitment, reduce biases, and increase the likelihood of reaping the economic, social and health benefits of a more diverse workforce? Rather than de-biasing minds through diversity leaders and training programs, research has shown that there are more fruitful opportunities to de-bias processes to encourage fairer recruitment practices. Dr. Iris Bohnet, a leading behavioral scientist in gender equality, argues that often people do not realize that they are biased and that they do not receive feedback in time to link a specific decision or behavior to bias. Bohnet recommends using behavioral design to level the playing field to create environments that make it easy to design fairer practices (Bohnet, 2016). A famous example of behavioral design is blind review, whereby any or all identification details of the candidates are removed when the rater reviews the candidate. One of the most famous examples of blind review is when orchestra directors held blind auditions by requiring musicians to play behind a curtain and then rated their performance. Although many of the orchestra directors reported not needing curtains because they only cared about the quality of music and not what somebody looked like, it turned out that curtains helped increase the proportion of women from about five percent in the 1970s to almost 40 percent in 2000 (Goldin & Rouse, 2000). This demonstrates the power of behavioral design because it helps people understand that sometimes small changes made in an environment, such as adding a curtain, makes it easier for well-meaning people to do the right thing. This experiment has

been successfully replicated in other settings such as resume screening in corporations (McIntyre et al., 1980) and entrepreneurial pitches (Brooks et al., 2014).

### Digital recruitment tools and biases

In addition to changes in recruitment processes, advancement in digital, specifically artificial intelligence (AI), shows promise in improving how organizations are overcoming unconscious bias in their hiring practices through technology. For example, after major companies like Unilever and L'Oréal partnered with AI hiring providers that used blind review, skills-based tests, and sophisticated content analysis software, they were able to double their application numbers, cut their review time by up to ninety (90) percent and dramatically increase the diversity of the candidate pool (Feloni, 2017; Sharma, 2019). At first glance, using AI-enabled recruitment tools seem to have mainly positive effects such as reducing time-to-hire and creating an opportunity for a more inclusive recruitment process. However, behavioral design and technology are not silver bullets. Recent evidence suggests that even under blind review disparities in outcomes exist across demographic groups, especially in the case of academia, research institutions, and funding opportunities (Shen, 2013; Witteman et al., 2018; Kolev et al., 2019). If the adoption of blind review cannot overcome (unconscious) biases, this raises the question of whether bias-driven mechanisms are the sole driver of demographic disparities. Differences in experience, risk tolerance, or context-specific structural elements may generate gaps across demographic groups even under a completely unbiased review process (Kelly et al., 2010; Murray & Graham, 2007; Kolev et al., 2019). More upstream structural drivers such as homogeneity of the educational pipeline or applicant pool may also contribute to demographic disparities (Shakhina et al., 2021).

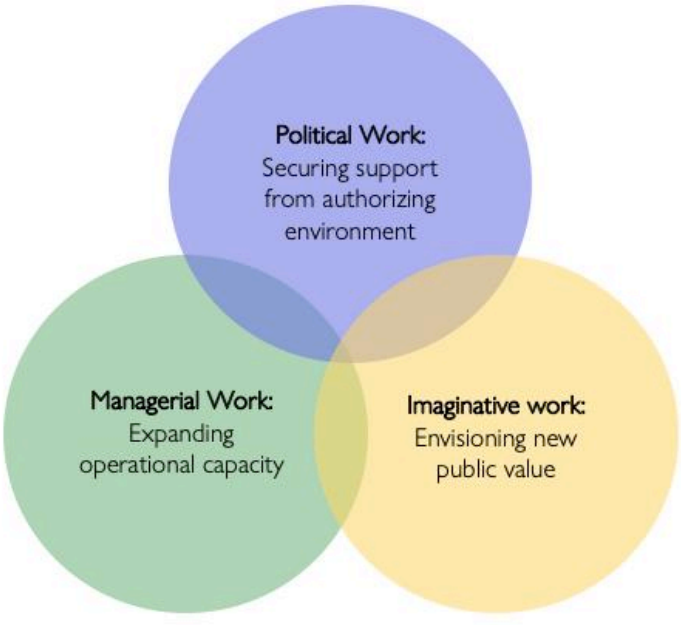
Technology also has severe limitations, especially when utilizing historical datasets to train algorithms. For example, Amazon had to get rid of its experimental AI-enabled recruiting

tool because it was trained on data that favored historically successful candidates (e.g., mostly white and male candidates) and resulted in a disadvantage to non-white and female candidates (Dastin, 2018). Without proactively testing for disparate impact in selection algorithms, practitioners may codify existing biases and fall prey to "data chauvinism" where they become overconfident in big data without ensuring that the data collection includes underprivileged groups (Lee & Viswanath, 2020). While organizations should aim to ensure their recruitment practices do not favor candidates based on non-merit considerations such as race or gender, blinded review or technology alone may not be sufficient if the goal is to eliminate the demographic disparities that have limited workforce diversity. In light of this discrepancy, before an organization adopts a new recruitment policy or tool it is worth evaluating past policies and proactively testing new strategies to understand what works to ensure fairer recruitment practices in a specific context.

## Chapter 3: Theory of Change

Learning from the behavioral science literature on recruitment bias solutions and how this has been applied in organizations, I aimed to find an opportunity where I could work with HR to measure what kind of bias (if any) exists in the Organization's recruitment process for the Project. A major challenge of the Project was to secure the resources and support necessary to approach the Organization's recruitment practices differently and ultimately, contribute to help building a more diverse and representative workforce. As my theory of change, I drew primarily from Mark Moore's "Strategic Triangle" – a three-part conceptual framework for innovation in the public sector that focuses on managing stakeholders, developing organizational capacity, and creating value for the organization (Moore, 2000). In 2012, Moore's framework was further adapted in the *Agents of Change* where the authors categorized this as three types of issues: political, managerial, and imaginative work (Figure 1) (Cels et al., 2012).

*Figure 1: Mark Moore's Strategic Triangle*



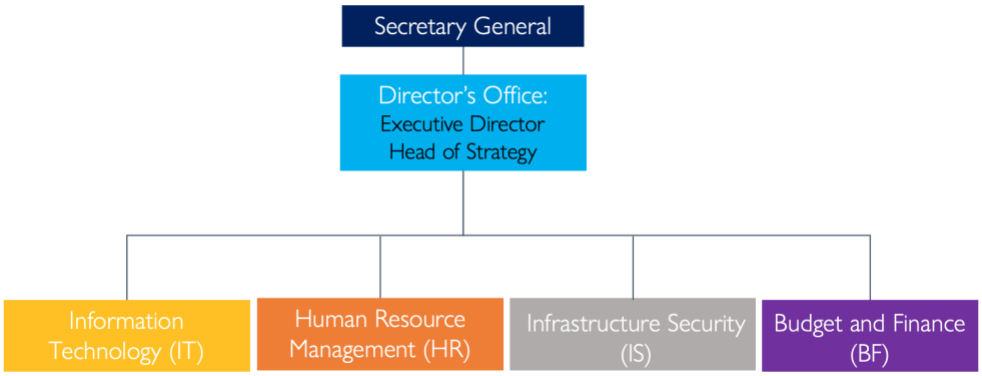

Leveraging the strategic triangle gave me the tools to diagnose the environment and its actors to inform the Project's design and delivery. This approach allowed me to reimagine what was

allowed, feasible, and valuable to the Organization and increase the Project's likelihood of success and adoption.

### Political Work

Enabling any kind of change requires obtaining and maintaining enough support inside the Organization to make the change urgent and necessary. According to Moore, the political work includes a diverse range of actors who exercise authority, enhance legitimacy, and give or withhold support from the authorizing environment (Moore, 2000b). The first step was to garner the support of the Director's Office (DO). The DO leads the Executive Department (ED) – the corporate arm of the Organization. In 2017, the new Head of Strategy of the DO received approval and funding from the Executive Director to pilot a behavioral design team with four graduate students for one year. The goal was to identify potential opportunities where behavioral science could be used to improve the design and delivery of corporate services. This included using a more data-driven approach and considering the potential unconscious biases that could affect staff's decision-making. Before the team's arrival, the Head of Strategy released a call for proposals to all the Services. Four services make up the ED and are directly under the leadership of the DO: Information technology (IT), human resources (HR), infrastructure security (IS), and budget and finance (BF) (Figure 2).

*Figure 2: Organigramme of the Executive Department (ED)*



The Service that expressed the most interest was HR, specifically the Deputy Head of HR. They already had plans to revisit their ethics policies and recruitment tools and saw this as an opportunity to leverage these resources. The Head of Strategy's ability to obtain approval from the Executive Director and identify potential early adopters outside of the DO were essential to getting the legitimacy necessary to look for potential areas of innovations inside the Organization.

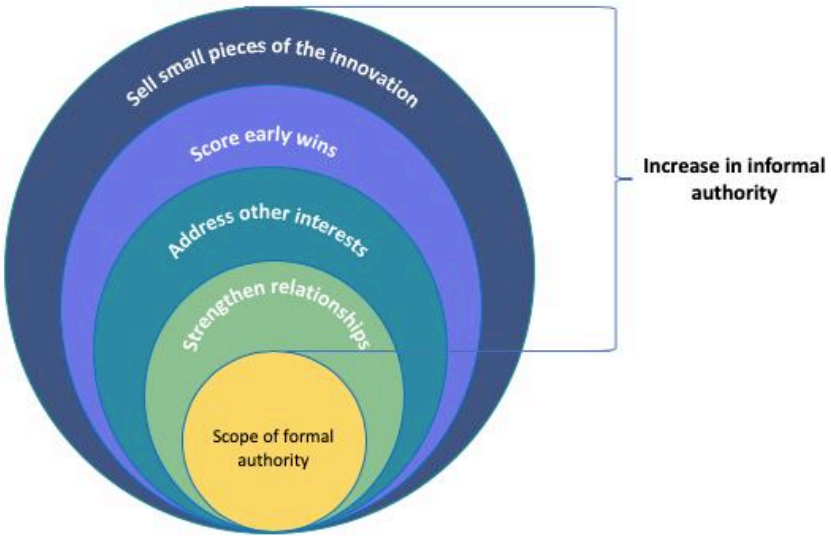
As part of the behavioral design team, my first month was full of meetings to inquire, listen, and observe the environment and its actors. These interactions allowed me to learn the lay of the land so when the time came to implement the Project so I could "act politically," which means:

> Using your awareness of the limits of your authority and of stakeholders interests as well as power and influence networks in your organization, to forge alliances with people who will support your efforts, to integrate and defuse opposition, and to give valuable dissenting voices a hearing. (Heifetz, 2009, page 132)

Learning how to act politically inside the Organization was necessary to maintain the legitimacy that the Head of Strategy initially secured. Navigating the authorizing environment informed the design of the Project and eventually mobilized the adaptive work.

I looked to Ronald Heifetz's two types of authorities to define the scope of the political work: informal and formal authority. Formal authority is the power given to a leader by an organization and depends on the power that comes from that authorizing environment. In contrast, informal authority is given to a leader by their followers (Heifetz, 1994). As a temporary staff member at the Organization, I had had neither formal nor informal authority. To gain any support from the Services, I expanded my informal authority through four phases (Figure 3).

*Figure 3: Phases to increase informal authority (Adapted form Heifetz, 2009)*




First, I focused on strengthening my relationships by forging connections with people who had a potential stake in using a more data-driven and behavioral science approach in their work. In the first month of meetings, I tried to gauge people's roles, priorities, and loyalties outside of formal meetings. It was only during the hallway conversations after the meetings or the impromptu encounters while waiting in line for coffee where I could learn what was important and necessary processes to make the work happen. These informal interactions helped me scan the environment for obstacles, concerns, and dissent. For example, instead of trying to directly convince the Deputy Head of HR, I learned that I needed find a frontline HR advisor get buy-in from the Deputy Head of HR for new projects. In my search to find the most influential HR advisor who would be interested in trying a behavioral science approach, I was able to identify who were the strong advocates of the status quo and who were keen for change.

Once I identified potential allies, the second step was to address their interests beyond the original objective. I was most successful when I shifted to listening to others' challenges instead of sharing my own. Some colleagues shared personal stories of their frustrations when their proposals were rejected, or they did not have the capacity to develop a new idea because

they were too busy keeping up with their daily tasks. For example, one of the HR advisors shared his eagerness to change the formal tone of communications to staff to identify ways to increase readership. Another advisor who was a data scientist in her previous job shared her concerns about the missed opportunities to measure diversity in their recruitment process. Our shared interests in finding ways to improve operations and generate better evidence helped create common ground.

The final phase involved scoring early wins by selling small pieces of a more data-driven and behavioral science approach. I found ways to bring value to smaller projects that were important to the ED rather than finding new projects to lead. For example, through my conversations with HR staff, I learned that improving internal ethical practices was a priority for the Deputy Head of HR. One of the challenges was to get staff to report to HR any expensive or electronic gifts that they received during business trips. At the Organization, receiving expensive gifts was seen to be problematic because it could hinder the impartiality staff. For example, receiving gifts can be a reciprocity mechanism where the person receiving the gift may feel obliged to give something in return such as preferential treatment (Ariely, 2012; Wazana, 2000). I saw this as an opportunity to build credibility with the Deputy Head of HR and have HR advisors get comfortable with the idea of using a behavioral science approach. I worked with HR staff to interview other employees and review relevant behavioral science research to understand the practical barriers and behavioral biases to gift reporting. Using these insights, we co-designed a gift reporting campaign that had two main objectives:

1. Increase the number of reported gifts received by the Organization's employees during December 2018-Feburary 2019.
2. Understand which motivation(s) is/are effective to enhance compliance of the gift reporting policy.

Appendix 1 provides a summary of the gift reporting campaign. Although the sample size was too small (n=4) to reach any conclusions on which message was the most effective to increase staff gift reporting, the behavioral science approach of collecting staff insights, reviewing relevant academic research, and evaluating the impact of the messages required staff to change their usual routine of sending the same annual staff email. The experiment proved to be a valuable proof-of-concept. In the end, going through a behavioral science approach with HR staff demystified the process, demonstrated the value-add, and generated interest from staff who wanted to try something different.

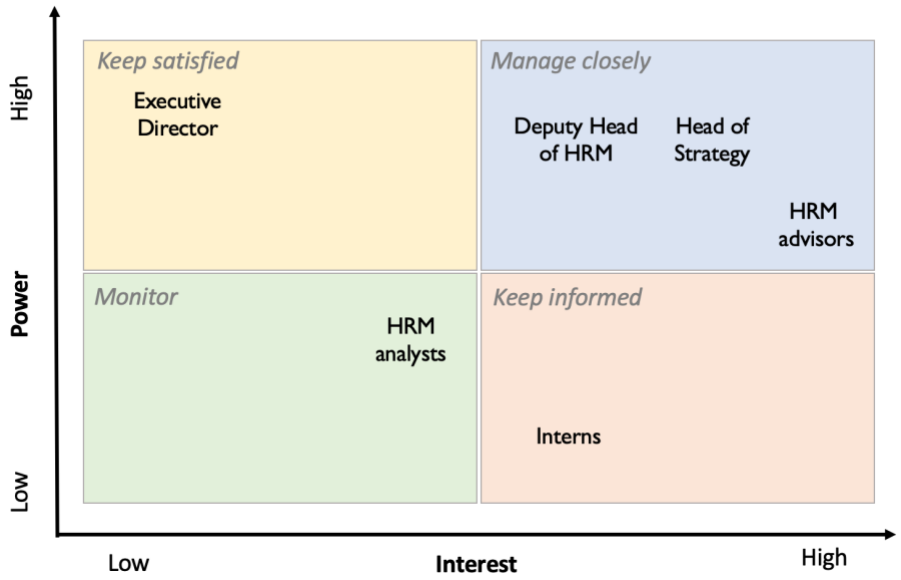
### Managerial Work

Once I started the political work of building a foundation of support from the authorizing environment, the next part involved the managerial work of securing sufficient operational capacity. According to the Strategic Triangle, operational capability includes sufficient resources correctly allocated, a skilled workforce, access to valuable data, and effective communication (Cels et al., 2012). A way to expand operational capacity is through assessing capacity and managing authority figures.

After reviewing our skills and goals as a behavioral design team, we assessed our current capacity and identified where we needed additional resources. As a team of four graduate students, we had the technical capacity to take on a behavioral science project with the right resources and data. However, we had limited operational ability to access the data, get senior and frontline buy-in, and create urgency to develop and implement a project. We first identified the senior stakeholders who were early adopters and had enough authority to remove barriers to obtain the data and advance the work. Then, we worked closely with their advisors who could navigate the day-to-day processes, and analysts who had data access and knowledge. Finally, we recruited interns who volunteered to commit a few hours a week to help with data collection and cleaning.

To prioritize the next steps, I mapped the stakeholders who would be impacted by the work and classified them according to their power (seniority in the Organization) and interest that was measured by a staff member's (lack of) response to an expression of interest form to receive behavioral science support for a corporate service challenge (Figure 4). The Executive Director had the highest seniority but the lowest level of interest. As the behavioral design team, we needed to keep her satisfied by informing her of our plan and outcomes relevant to the broader organizational strategy. However, we were careful not to overwhelm her with continuous updates and tried to stay under her radar. Since we already built relationships with two key HR advisors from the gift reporting campaign, we spent two weeks identifying priority challenges that would benefit from a behavioral science approach. With the two HR advisors we were able to identify a project evaluating the recruitment software used to record job applicants' video interviews. Co-designing the projects with HR enabled us to access additional resources because they were the gatekeepers to the analysts and interns who could ensure that the work could advance on time. Finally, we identified that the Head of Strategy and Deputy Head of HR had significant seniority in the Organization and expressed deep interest in integrating a more behavioral science approach, so we engaged them throughout the process and updated them at key milestones.

*Figure 4: Stakeholder Map*




## Imaginative Work

The work involved capacity beyond the behavioral design team. It required convincing each stakeholder group that their efforts would be for a worthwhile and significant cause that mattered to them. Moore calls this "crafting a public value proposition" in such a way that it becomes not just a novel idea but a feasible solution to a problem that addresses an organizational challenge (Moore, 2000). Crafting the value proposition involved explicitly outlining and agreeing on the performance and outcome metrics that are of importance to stakeholders. In *Agents of Change*, Cels et. al., recommend the need for "aggressively managing meaning" because of the following:

> There is no such thing as one fixed meaning of an innovation; stakeholders may each make sense of its public value in their own way. Some may share the views of the original innovator; some may see different value in it from another perspective; and others may appreciate the project but hesitate to support it because they think it does not fit into the mission of their organization. Innovators need to manage the process of sense making: they need to guide stakeholders toward a favorable interpretation (and course of action) (Cels et al., 2012).

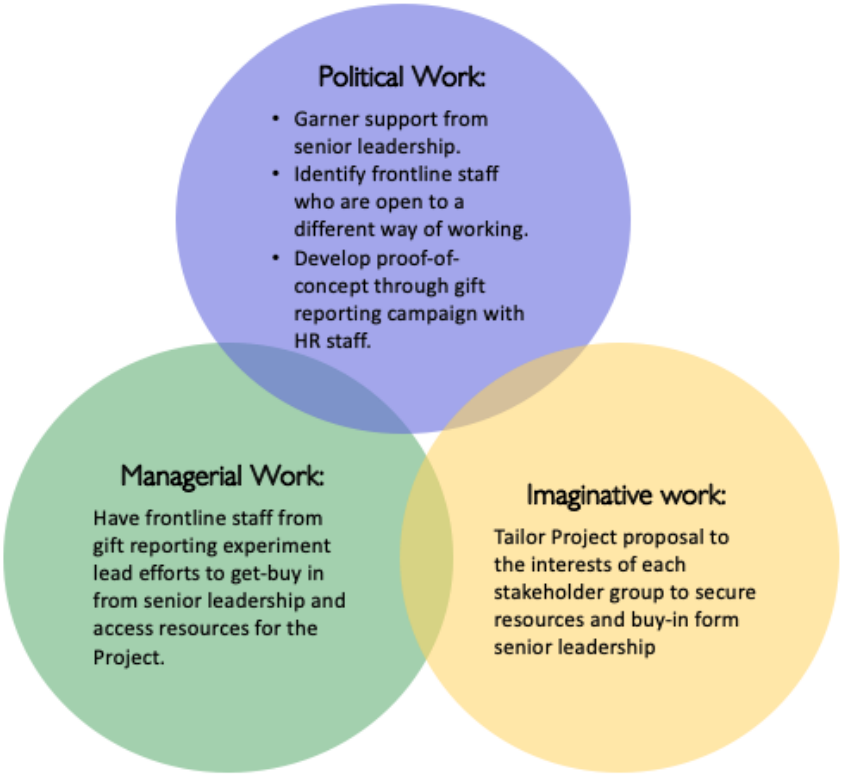
Being part of hallway conversations and learning the organization's corporate priorities allowed me to "manage meaning" and understand the loyalties, vision, and values of key stakeholders. From there, I used these insights to paint different pictures of success, depending

on what was relevant for each person. For example, in my conversations with the Deputy Head of HR, I focused on the financial benefits of evaluating the digital recruitment tool. I provided other examples of organizations that had contained costs by introducing new technology rather than looking at the processes and procedures that contributed to the problem. In some cases, the latest technology had created its own set of challenges and did not produce direct cost savings. This argument resonated with her because she had an upcoming budget review, which included decisions around contract renewal with the provider. With the HR advisors, I focused on the opportunity to use this work to prepare for next year's gender equality re-certification. They were motivated to do work related to this topic because the Organization missed the gender equality certification by a few points in previous years. Finally, with the Head of Strategy, who wanted to build a sustainable behavioral science portfolio within the Executive Directorate (ED), I framed these projects as a new way of working that allowed for knowledge and skills transfer to ED staff. The capacity-building perspective aligned with his vision of mainstreaming behavioral science within the Organization by having non-behavioral science experts gradually initiate behavioral science work themselves and still work with the trained experts where it was necessary. By "managing meaning" in a tailored way to each stakeholder group, I was able to modify and reimagine the work in terms of what was valuable and feasible to implement.

## Conclusion

By leveraging Moore's Strategic Triangle, I was able to manage stakeholders, identify organizational capacity, and create a vision that brought value to the organization. Figure 5 provides an overview of how the Project was secured using Moore's Strategic Triangle.

*Figure 5: Summary of how the Project was secured using the Moore's Strategic Triangle framework*




Using this approach allowed me to secure a project with HR reviewing potential biases in their new digital recruitment tool. The next chapter describes the background, inherited study design, and analysis of the Project.

## Chapter 4: Analysis
## Introduction

### Background: Recruiting young talent at the Organization

Since the 1970s, the Organization has recruited talented young graduates who show great potential for leadership positions through the YP program (YPP). Every two years, directorates across Organization request a YP to serve as an entry-level policy analyst whose technical expertise aligns with the strategic orientation of the Secretary-General. YPs are typically master's or PhD level graduates who continue onto managerial roles at the Organization. More recently, the Organization developed the YA program (YAP) to target final-year or recent undergraduates. Compared to YPs, YAs are less experienced and use the 24-month program to build the necessary professional competences and policy expertise needed to secure places in highly competitive master's and doctoral programs in the future. In previous years (2010-2018), HR recruitment evaluators ("raters") first screened CVs of YPP and YAP candidates ("candidates"). For the second stage, raters sent invites to the top 10% (typically 60-80 candidates) to participate in a phone video interview. From there, successful candidates participated in a written test, second phone interview, and a panel interview. Figure 6 outlines each stage of the typical recruitment process.

*Figure 6: Five stages of 2010-2018 recruitment process*



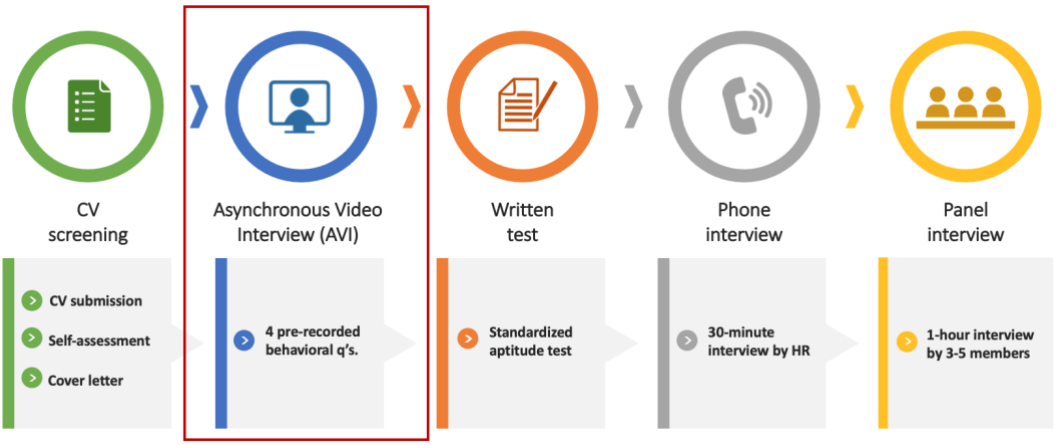
## Rationale

### *Rationale to pilot asynchronous video interviews (AVIs)*

During the 2018-19 YPP and YAP recruitment process, the Organization had the opportunity to trial a recruitment software that used asynchronous video interviews (AVIs). The AVI software sent each candidate a link to a website where they were asked four behavioral questions and had one minute to answer each question. Their answers were recorded and sent back to the raters to review and score each answer. In contrast to "two-way" synchronous video interviews such as Skype, AVIs are known as "one-way" interviews whereby candidates are prompted with pre-recorded questions. Their answers are recorded at one point in time and evaluated at another point in time (Mejia & Torres, 2018). AVIs have become increasingly popular and have replaced traditional phone or conference interviews because they are more time- and cost-efficient (Suen et al., 2019). AVIs are an attractive recruitment tool for international organizations because they have candidates from all around the world. With AVIs, HR professionals can evaluate the interviews on their own time and not waste time trying to find an interview time that is appropriate for multiple time zones.

For the 2018-19 YPP and YAP recruitment process, raters used AVIs to replace the first phone interview during the second stage of the recruitment process (Figure 7). Since AVIs allowed for more time-efficient recruitment, senior HR advisors decided to invite the top 30% (instead of the usual 10%) of the 986 candidates from the CV screening stage. At the AVI stage, 296 candidates were reviewed for 12 positions. Out of the 296 candidates, those who scored in the top 30% (80 candidates) moved on to the written test.

*Figure 7: Five stages of 2018-2019 recruitment process*




### *Rationale to analyze different mediums*

The Organization's trial with the AVI supplier was coming to an end. HR, specifically the Deputy Head of HR and senior HR advisors, had to decide whether to use AVIs as part of their routine recruitment practice. Although the AVI software allowed them to review more candidates compared to phone interviews in previous years, a concern that was raised by HR was whether the video medium introduced a bias towards a specific type of candidate based on non-merit considerations, specifically gender and nationality. They wanted to make an informed decision on whether to continue with AVIs or seek other software that only used audio interviews or transcribed interviews. To inform their decision, HR implemented a pilot where they converted the AVIs to audio and transcript versions for the raters to reviews in addition to the AVIs. The Methods section outlines the pilot (inherited study design) in more detail. After the pilot was implemented, HR sought the support of the behavioral design team to help analyze the results.

## Objectives

Initially, HR sought the support of the behavioral design team to analyze the video, audio, and transcript versions of the AVIs to test whether there was a bias towards candidates

from a specific nationality or gender. Unfortunately, since the pilot was implemented before a detailed analysis plan was created with the behavioral design team, HR and the behavioral design team had to redefine the objectives within the scope of the inherited study design. After discussions with HR, we focused on conducting an analysis to test the extent to which non-merit considerations, specifically gender and nationality of the candidate, and rater's characteristics (gender and nationality of rater) biased the mean interview score. Additionally, we focused on understanding if different interview mediums influenced how HR staff evaluated YAP and YPP candidates. In the end, the objectives of the overall AVI pilot were:

1) To select YAP and YPP candidates for the written test stage.
2) To determine if there was an association between non-merit demographic characteristics (specifically, the candidate's gender, candidate's nationality, rater's gender, and rater's nationality) and the mean interview score evaluated by HR recruiters ("raters").
3) To determine if there was an association between similar demographics (same gender or same nationality between rater and candidate) and the mean interview score evaluated by raters.
4) To determine if there was an association between different interview mediums (video, audio, or transcript interviews) and the mean interview score evaluated by raters.

The main objective of the Project was to analytically explore objectives 2 -4.

### Hypotheses

To achieve the objectives 2-4, I tested four hypotheses that were co-developed in discussions with the Deputy Head of HR, HR advisors, and the behavioral design team. Since each candidate had a video, audio, and transcript version of their interviews the unit of analysis was an interview and not a candidate.

### *Hypothesis 1: Demographic characteristics*

The first hypothesis assumes that surface-level, non-merit considerations such as demographic characteristics of the candidates will influence how the interview is evaluated. Studies have consistently shown that despite an organization's efforts to develop a more merit-based recruitment process and intention to build a more diverse workforce, too often there is still a negative bias towards women and people from minority backgrounds (Isaac et al., 2009; Roper, 2019; Vick & Cunningham, 2018). Building on these findings, I hypothesize that there will be a significant association between the mean interview score and the following demographic characteristic: female candidate, candidate from top Member country, and female rater. Specifically:

H1a: Interviews with male candidates will be scored hired than interviews with female candidates.

H1b: Interviews with candidates from the top five highly represented Member countries (France, Germany, Italy, United States and United Kingdom) will be scored hired than interviews with candidates from other Member countries.

H1c: Interviews with male raters will be scored higher than interviews with female raters because previous studies have shown that women tend to be harsher evaluators (Furnham & Stringfield, 2001; Payne, 2011).

H1d: Interviews with raters from the top five highly represented countries will not significantly differ from interviews with raters from other countries because previous studies, although limited, have found little difference between raters from majority and minority groups (Sackett & DuBois, 1999.; Shore et al., 2009).

I used the following specifications for H1:

$$mean\ interview\ score = \alpha + \beta_1\ female\ candidate + \beta_2 candidate\ from\ top\ Member\ country + \beta_3 f\ emale\ rater + \beta_4 rater\ from\ top\ Member\ country + \epsilon$$

α is the intercept, $\beta_1$ is the main effect of a female candidate, $\beta_2$ is the main effect of candidate from top Member country (France, Germany, Italy, United Kingdom, United States), $\beta_3$ is the main effect of female rater, $\beta_4$ is the main effect of rater from top Member country (France, Germany, Italy, United Kingdom, United States), and ε is an error term.

*Hypothesis 2 (H2) and Hypothesis 3 (H3): Similarity-attraction paradigm*

The second and third hypotheses are derived from the demographic similarity theory, which suggests that people can have a preference towards others who are similar to them or share group-level characteristics, such as gender and nationality (Lawrence & Shah, 2020; McCarthy et al., 2010). This could be explained by the similarity-attraction paradigm, which theorizes that individuals regard others more positively when they are viewed as more similar to themselves because it is assumed that individuals with similar demographics will also have similar underlying attributes (Milliken & Martins, 1996). In the interview context, the potential for similarity-attraction paradigm can have serious implications as it may result in unfavorable selection decisions for dissimilar applicants. As a result, there may be a bias towards candidates who share similar either the same gender or same nationality with the rater and ultimately contribute to a homogenous workforce. To test the similarity-attraction paradigm in the Organization's recruitment context, I first explored the extent to which interviews where having the same gender between the candidate and rater affected the mean interview score.

H2) Interviews where raters share the same gender with the candidate will result in higher mean interview scores compared to interviews where raters do not share the same gender. For H2, I used the following specifications:

Model 2:

$$mean\ interview\ score$$
$$= \alpha + \beta_1 female\ candidate + \beta_2 candidate\ from\ top\ Member\ country$$
$$+ \beta_3\ female\ rater + \beta_4 rater\ from\ top\ Member\ country$$
$$+ \beta_5\ female\ candidate * female\ rater + \epsilon$$

α is the intercept, $\beta_1$ is the main effect of a female candidate, $\beta_2$ is the main effect of a candidate from a top Member country (France, Germany, Italy, United Kingdom, United States), $\beta_3$ is the main effect of a female rater, $\beta_4$ is the main effect of a rater from a top Member country (France, Germany, Italy, United Kingdom, United States), $\beta_5$ is the interaction effect of candidate's gender and rater's gender, and ε is an error term.

Additionally, I explored the extent to which interviews where having the same nationality between the candidate and rater affected the mean interview score.

H3) Interviews where raters share the same nationality with the candidate will result in higher mean interview scores compared to interviews where raters do not share the same nationality. For H3, I used the following specifications:

Model 3:

$$mean\ interview\ score$$
$$= \alpha + \beta_1 female\ candidate + \beta_2 candidate\ from\ top\ Member\ country$$
$$+ \beta_3 female\ rater + \beta_4 rater\ from\ top\ Member\ country$$
$$+ \beta_5\ candidate\ from\ top\ Member\ country$$
$$* rater\ from\ top\ Member\ country + \epsilon$$

α is the intercept, $\beta_1$ is the main effect of a female candidate $\beta_2$ is the main effect of a candidate from a top Member country (France, Germany, Italy, United Kingdom, United States), $\beta_3$ is the main effect of a female rater $\beta_4$ is the main effect of a rater from a top Member country (France, Germany, Italy, United Kingdom, United States), $\beta_5$ is the interaction effect of a

candidate from a top Member country (France, Germany, Italy, United Kingdom, United States) and a rater from a top Member country (France, Germany, Italy, United Kingdom, United States).

### *Hypothesis 4 (H4): Media richness*

The fourth hypothesis is derived from media richness theory, which suggests that richer mediums such as video interviews that provide visual, emotional, or behavioral cues can reduce the risk of miscommunication and have a strong explanatory power in the selection process of job candidates compared to audio or transcript interviews (Blackman, 2002). Based on the media richness theory, I hypothesized that raters would evaluate transcript interviews less favorably than video or audio interviews because they would not be able to access more social cues.

H4) Interviews in higher media richness will result in higher average scores. Specifically, video and audio interviews will be scored higher than transcript interviews for each candidate. For H4, I used the following specifications:

Model 4: $mean\ interview\ score = \alpha + \beta_1 medium + \epsilon$

where α is the intercept, $\beta_1$ is the main effect of the medium and ε is an error term

## Methods

### *Inherited study design (pilot)*

For the 2018-19 YAP and YPP recruitment process, the top 30% of the 986 candidates from the CV screening stage were invited to participate in an interview as part of the second recruitment stage (AVIs). Using the AVI software, each candidate was asked an interview question on their motivation, overcoming a challenge, relevant experience, and views on multilateral organizations (four questions total). Details of the questions and response criteria are found in Appendix 3. After all the candidates submitted their AVIs, a HR professional who was not part of the hiring committee converted the video interviews to audio files and

transcribed them to scripts using a transcription service. In the end, each candidate had video, audio, and transcript versions of their interview.

Sixteen (16) raters made up the hiring committee that participated in the evaluation. For every job vacancy, raters were grouped into teams of four, which met the following criteria:

1. At least one male rater
2. Every rater holds a different nationality
3. Evaluated at least one YP and YA job vacancy

Each rater independently evaluated their assigned interviews and were not influenced by the other raters in their group. Raters were split into teams of at least three and were assigned to three to four job vacancies out of the 12 total job vacancies listed in Appendix 2. Since a rater may remember a candidate if they reviewed more than one medium per candidate, raters scored each candidate in only one of the mediums. Thus, all three mediums of each candidate were independently reviewed by a different rater. Table 1 illustrates the design.

*Table 1: Inherited study design*

| | Rater 1 | Rater 2 | Rater 3 |
|---|---|---|---|
| Candidate 1 | Video | Audio | Transcript |
| Candidate 2 | Audio | Transcript | Video |
| Candidate 3 | Transcript | Video | Audio |
| … | … | … | … |
| Candidate n | Video | Transcript | Audio |

Every rater evaluated the interviews in the following order: video, audio, and transcript. All raters were coached on criteria and standardization of questions and evaluations prior to scoring the interviews, which is standard practice in the Organization because structured interviews where each candidate is asked the same questions in a predetermined order and responses are graded using a pre-specified standardized criteria has been shown to reduce the impact of bias (Levashina et al., 2014). Each rater provided a score for the candidate's response to each of the four interview questions on a scale from 1 to 5. All four questions were weighted

equally. After the raters finished evaluating the assigned interviews, they submitted their scores to an HR professional who was not part of the hiring committee. Finally, the HR professional who was not part of the hiring committee calculated the mean interview score for each candidate based on the average rating across all four questions for each medium (video, audio, and transcript).

### *Variables of interest to the Organization*

The following section defines the candidate variables, rater variables, main exposure variables, and outcomes measures that were used in the pilot analysis.

#### Candidate variables

##### Female candidate

At the time of the application process, candidates were asked to report their gender and chose among "male," "female," and "not specified." Since all 254 candidates selected either "male" or "female," a binary variable was used to represent gender (1=female, 0=male).

##### Candidate from a top Member country

The Organization hires candidates who hold at least one nationality from one of its member countries (Members). Members are countries that have successfully undergone a complex review process and demonstrated a commitment to protection of human rights and free-market economies. At the time of application, candidates were required to report their nationality. While there are many ways to measure nationality, a binary variable was selected to identify those nations with a higher rate of representation overall. The top five nations were chosen as an organizationally relevant threshold by senior HR advisors for this analysis based on two criteria. First, when HR reviewed the annual staff member composition over the past few years, they reported that while the second, third, and fourth places may interchange among Members, the top five Members were relatively consistent and predictable. Second, HR reported that the difference in representation rates was more than one percentage point between

the 5th and 6th ranks in 2018, showing at least a judgmental gap between the top five and the other members (i.e., Members with greater than or equal to the 6th rank generally had low differences and a grouping there could appear quite arbitrary and subject to small changes in new staff year to year). Therefore, the top five grouping seemed the least sensitive to potential small changes in staff over time and a meaningful grouping to understand the summary representation instead of looking at each of the Members individually as dummy variables, which would be problematic on such a small sample size of a few hundred. The top five nations include: France, Germany, Italy, United Kingdom, and United States. Thus, a dummy variable was used for the nationality of a candidate (1=holds nationality from France, Germany, Italy, United Kingdom, and United States, 0=holds other nationality).

### Rater variables

#### Female rater

For the purposes of the study, raters were asked to report their gender. Since all 16 raters selected either "male" or "female," binary variable was used to represent gender (1=female, 0=male).

#### Rater from a top Member country

At the start of the study, raters were also asked to report their nationality. The same top five grouping that was used for candidates was used for raters to easily compare between candidates and raters. Thus, the nationality of raters was also a dummy variable (1=holds nationality from France, Germany, Italy, United Kingdom, and United States, 0=holds other nationality).

### Main exposure variable

Medium

"Medium" represented the mode in which the interview was evaluated by the rater. Every candidate had a score for each medium: video, audio, or transcript. Medium was defined as a categorical variable where transcript was the reference level (0=transcript, 1=audio, 2=video).

### Outcome variable

Mean interview score

For each interview, the mean interview score was calculated by taking the sum of each interview score provided by the rater and dividing by four (total number of questions asked per interview). For each interview question, a rater gave a score on a scale from one (poor performance) to five (good performance). Appendix 3 lists each question and evaluation criteria in detail. All interview questions were weighted equally.

### Statistical Analysis

#### *Bivariate analysis*

I conducted a bivariate analysis in JMP software to explore the association between the mean interview score (outcome) and similar demographics variable. I ran a one-way ANOVA to understand the association between the medium and mean interview score. I used Tukey's HSD Test for multiple comparisons to compare the difference in mean interview score between video and transcript interviews, audio and transcript interviews, and video and audio interviews.

#### *Linear regression analysis*

I then ran linear regressions for each hypothesis to quantify the association between the mean interview score and predictors in all models, using an alpha level of 5% (two-tailed) to determine statistical significance. For Model 4, I tested a model with each potential confounder (adjusted model) against the crude model (effect of medium on outcome) using the 10% rule

for confounding: If the percent change of the beta estimate of the main exposure between the crude model (effect of medium on the mean interview score) and adjusted model (effect of medium on the mean interview score controlling for a hypothetical confounding variable) was greater than 10%, the variable was considered to be a confounder (VanderWeele, 2019).[1] Hypothetical confounders included only demographic characteristics: gender of candidate, gender of rater, nationality of candidate, and nationality of rater.

Once I identified confounders using the 10% rule for confounding, I compared different adjusted models using Akaike's Information Criteria (AIC) and Bayesian Information Criteria (BIC). The AIC is commonly used to measure the goodness of fit of any estimated statistical model. The BIC is a type of model selection among a class of parametric models with different numbers of parameters. The model with the lowest AIC/BIC was the final adjusted model.

## Results

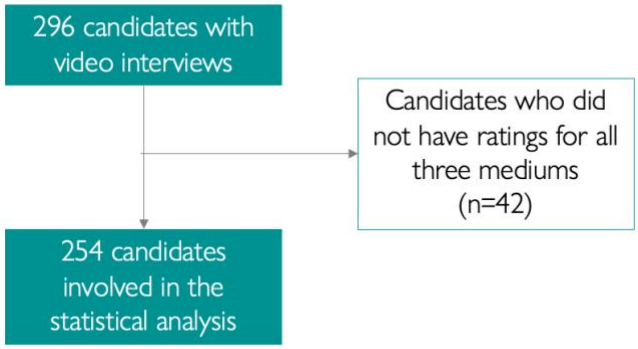
### Participants

The initial data set from the inherited study design had a total of 296 candidates. For the purposes of the analysis, only candidates who had ratings for all three mediums (video, audio, and transcript) at least once were included in the dataset. This included candidates who had multiple ratings for one medium (i.e., two video, one audio, and three transcript ratings). Candidates who did not have all mediums were excluded (n=42) because there may have been cases where the candidates who had only video scores were inherently different from candidates who had all mediums. This could inadvertently appear as showing higher scores, thus better qualified, in the video interviews even if the cause may have been unrelated (i.e., corrupt audio file). Based on these inclusion criteria, 42 candidates were excluded but had

[1] $magnitude\ of\ confounding\ (\%) = \left[\frac{(beta\ estimate\ of\ the\ main\ exposure\ in\ the\ crude\ model - beta\ estimate\ of\ the\ main\ exposure\ in\ the\ adjusted\ model)}{beta\ estimate\ of\ the\ main\ exposure\ in\ the\ crude\ model}\right] * 100$

similar demographic characteristics compared to the candidates who were included. In the end, the dataset had 254 candidates and an official response or inclusion rate of 0.86 (Figure 8).

*Figure 8: Flow chart showing inclusion criteria of data sample*




Among the included candidates, there was a slight female majority (54.33%) and most did not hold nationalities from France (FR), Germany (DE), Italy (IT), United Kingdom (UK), or United States (US). Among the 16 raters who were part of the hiring committee, many raters were women and, similarly to the candidate pool, most did not hold nationalities from FR, DE, IT, UK, or US. The demographic characteristics of included candidates and raters of the hiring committee are shown in Table 2.

*Table 2: Demographic characteristics of all included candidates and raters of the hiring committee*

| | Candidates | | Raters | |
|---|---|---|---|---|
| | n | % | n | % |
| **Gender** | | | | |
| Female | 138 | 54.33 | 10 | 62.50 |
| Male | 116 | 45.67 | 6 | 37.50 |
| **Nationality** | | | | |
| FR,DE, IT, UK, or US | 68 | 26.77 | 5 | 31.25 |
| Other | 186 | 73.22 | 11 | 68.75 |

## Descriptive data

A total of 974 interviews from included candidates were analyzed for the study. Roughly, there was an even distribution among the types of medium (transcript, audio, and

video interviews). Interviews where rater and candidate shared the same gender represented almost half of the sample (49.18%), but there were only a few instances where the rater and candidate shared the same nationality grouping (2.98%) (Table 3).

*Table 3: Interview characteristics by medium and shared demographics*

| Characteristic | n | % |
|---|---|---|
| **Medium** | | |
| Transcript interviews | 324 | 33.26 |
| Audio interviews | 325 | 33.37 |
| Video interviews | 325 | 33.37 |
| Same gender between rater and candidate | 495 | 49.18 |
| Same nationality between rater and candidate | 29 | 2.98 |

## Main results

### *Bivariate analysis results*

A one-way ANOVA analysis revealed that there was a statistically significant difference between the medium and mean interview score ($F(2, 970) = 4.48$, $p = 0.01$). Upon further exploration using Tukey's HSD Test for multiple comparisons, there was a significant difference between video and transcript interviews ($p = 0.01$, 95% C.I. = 0.03, 0.31). The difference in mean interview score between audio and transcriptive interviews was not significant ($p=0.05$). Similarly, the difference between audio and transcriptive interviews was not significant ($p=0.88$).

### *Linear regression results*

#### Model 1 results: Demographic characteristics

For H1, a multivariable linear regression was calculated to determine the association between the mean interview score and demographic characteristics (candidate's gender, candidate's nationality, rater's gender, and rater's nationality) (Table 3, Model 1).

*Table 4: Linear regression outputs for Models 1-4*

| | Model 1 [2] | | | | Model 2 [2] | | | | Model 3 [2] | | | | Model 4 [2] | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Predictors | α[1] | β [3] | 95% CI | p-value | α[1] | β [3] | 95% CI | p-value | α[1] | β [3] | 95% CI | p-value | α[1] | β [3] | 95% CI | p-value |
| Female candidate | 2.93 | **-0.10** | **-0.20, -0.02** | **0.02** | 2.95 | -0.10 | -0.30, 0.01 | 0.08 | 2.93 | **-0.11** | **-0.20, -0.02** | **0.02** | 3.20 | — | — | — |
| Candidate from top Member country[4] | | **0.39** | **0.29, 0.49** | **<0.01** | | **0.39** | **-0.29, 0.49** | **<0.01** | | **0.40** | **0.28, 0.52** | **<0.01** | | — | — | — |
| Female rater | | **0.52** | **0.43, 0.62** | **<0.01** | | **0.50** | **0.36, 0.64** | **<0.01** | | **0.52** | **0.43, 0.62** | **<0.01** | | — | — | — |
| Female from top Member country[4] | | -0.1 | -0.18, 0.01 | 0.08 | | -0.1 | 0.19, 0.01 | 0.08 | | -0.08 | -0.19, 0.03 | 0.16 | | — | — | — |
| Female candidate * Female rater | | — | — | — | | 0.05 | -0.14, 0.24 | 0.63 | | — | — | — | | — | — | — |
| Candidate from top Member country[4] * Rater from top Member country[4] | | — | — | — | | — | — | — | | -0.02 | -0.24, 0.19 | 0.82 | | — | — | — |
| *Medium* | | | | | | | | | | | | | | | | |
| Video vs transcript | | — | — | — | | — | — | — | | — | — | — | | **0.1** | **0.02, 0.26** | **0.02** |
| Audio vs transcript | | — | — | — | | — | — | — | | — | — | — | | **0.2** | **0.05, 0.29** | **0.01** |

[1] α= intercept
[2] See Table 5 for Legend
[3] β = regression coefficient (beta)
**BOLD** = Statistically significant (alpha=0.05)
[4]France, Germany, Italy, United Kingdom, United States

*Table 5: Legend that describes Models 1-4*

| | |
|---|---|
| Model 1 | $$\text{mean interview score} = \alpha + \beta_1\, female\ candidate + \beta_2 candidate\ from\ top\ Member\ country\ + \beta_3\, female\ rater + \beta_4 rater\ from\ top\ Member\ country + \epsilon$$ |
| Model 2 | $$\text{mean interview score} = \alpha + \beta_1 female\ candidate + \beta_2 candidate\ from\ top\ Member\ country + \beta_3\, female\ rater + \beta_4 rater\ from\ top\ Member\ country + \beta_5\, female\ candidate * female\ rater + \epsilon$$ |
| Model 3 | $$\text{mean interview score} = \alpha + \beta_1 female\ candidate + \beta_2 candidate\ from\ top\ Member\ country + \beta_3 female\ rater + \beta_4 rater\ from\ top\ Member\ country + \beta_5\, candidate\ from\ top\ Member\ country * rater\ from\ top\ Member\ country + \epsilon$$ |
| Model 4 | $$\text{mean interview score} = \alpha + \beta_1 medium + \epsilon$$ |

Model 1 resulted in the following specifications:

*Mean interview score = 2.93 -0.10(female candidate) + 0.39 (candidate from top Member country) +0.52 (female rater) +$\epsilon$*

Compared to all other interviews that had a mean interview score of 2.93,

a) interviews with female candidates were scored on average 0.10 points lower ($p=0.02$).
b) interviews with candidates from the top five highly represented countries (France, Germany, Italy, United States, United Kingdom) were scored on average 0.39 point higher ($p<0.01$).
c) interviews with female raters were scored on average 0.52 points higher ($p<0.01$).
d) interviews with raters from the top five highly represented countries were not scored significantly higher/lower because the association between the rater's nationality and the mean interview score did not reach statistical significance ($p=0.08$).

Overall, the results indicated that interviews with candidates who were male and interviews with candidates from France, Germany, Italy, United States, or United Kingdom were generally more favored. Model 1 results also revealed that female raters were on average more generous than male raters. Thus, these results supported hypotheses H1a, H1b, and H1d but did not support H1c.

### Model 2 results: Similarity-attraction paradigm (female candidate and female rater)

For H2, I hypothesized that interviews where raters share the same gender with the candidate will result in higher mean interview scores compared to interviews where raters do not share the same gender. A multivariable regression was conducted to determine the association between the interaction effect of candidate's gender and rater's gender and the mean interview score, controlling for demographic characteristics (female candidate, candidate from a top Member country, female rater, rater from a top Member country). The association between the interaction effect of female candidate and female rater and the mean interview score did not reach statistical significance (p=0.63) (Table 3, Model 2). Thus, interviews where the rater shared the same gender with the candidate did not significantly give the respective candidates higher interview scores compared to raters who did not share the same gender with the candidate.

### Model 3 results: Similarity-attraction paradigm (candidate and rater from top Member country)

For H3, I hypothesized that interviews where raters share the same nationality with the candidate will result in higher mean interview scores compared to interviews where raters do not share the same gender. A multivariable regression was conducted to determine the association between the interaction effect of a candidate from a top Member country and rater from a top Member country and mean interview score, controlling for demographic characteristics (female candidate, candidate from a top Member country, female rater, rater from a top Member country). The association between the interaction effect of candidate from a top Member country and rater from a top Member country score did not reach statistical significance (p=0.82) (Table 3, Model 3). Thus, interviews where both the rater and candidate were from a top Member country did not significantly give the respective candidates higher interview scores compared to raters who did not share the same nationality group (top Member country) with the candidate.

### Model 4 results: Media richness

For H4, I hypothesized that interviews in higher media richness will result in higher average scores. A simple linear regression was conducted to determine the association between mean interview score and medium (Table 3, Model 4). The association between medium and mean interview reached statistical significance ($p<0.01$) for Model 4, which is depicted in the equation below where transcript is the reference level for medium:

$$\textit{Mean interview score} = 3.18 + 0.14\ (audio) + 0.17\ (video) + \varepsilon$$

Audio interviews were scored 0.141 points higher than transcript interviews (p=0.02). Video interviews were scored 0.169 points higher than transcript interview (p=0.01). These results supported the hypothesis that transcript interviews would be scored the lowest. No confounders were identified for this model using the 10% confounding rule. Hence, there was

no adjusted model. Overall, raters scored candidates higher when they used video or audio mediums compared to transcripts, suggesting that richer mediums (video and audio) were generally favored by raters.

## Additional analysis

### *Stratified analysis*

To explore whether Model 1 findings were consistent across different medium (transcript, audio, and video), I ran the following specifications and stratified by medium:

$$mean\ interview\ score = \alpha + \beta_1\ female\ candidate + \beta_2 candidate\ from\ top\ Member\ country + \beta_3 f\ emale\ rater + \beta_4 rater\ from\ top\ Member\ country + \epsilon$$

α is the intercept, $\beta_1$is the main effect of a female candidate, $\beta_2$ is the main effect of candidate from top Member country (France, Germany, Italy, United Kingdom, United States), $\beta_3$ is the main effect of female rater, $\beta_4$ is the main effect of rater from top Member country (France, Germany, Italy, United Kingdom, United States), and ε is an error term. For transcript interviews, results were consistent with overall Model 1 findings (Table 6):

- Interviews with female candidates were scored on average 0.15 points lower (p=0.04).
- Interview with candidates from the top five highly represented countries were scored on average 0.49 points higher (p<0.01).
- Interviews with female raters were scored on average 0.51 points higher (p<0.01).

For audio and interviews, however, the association between candidate's gender and mean interview score did not reach statistical significance (Table 6). Thus, the stratified analysis showed that the relationship between the candidate's gender and mean interview score

was less pronounced in richer mediums (audio and video interviews) compared to transcript interviews.

*Table 6: Linear regression output to assess if the association between demographics (candidate's gender, candidate's nationality, rater's gender, and rater's nationality) and mean interview score differed by medium subgroup (transcript, audio, and video)*

| | Transcript | | | | Audio | | | | Video | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Predictors | α[1] | β[2] | 95% CI | p-value | α[1] | β[2] | 95% CI | p-value | α[1] | β[2] | 95% CI | p-value |
| Female candidate | 2.80 | **-0.15** | **0.31, 0.01** | **0.04** | 2.93 | -0.07 | -0.30, 0.01 | 0.34 | 3.07 | -0.12 | -0.29, 0.05 | 0.15 |
| Candidate from top Member country* | | **0.49** | **0.32, 0.65** | **<0.01** | | **0.36** | **-0.29, 0.49** | **<0.01** | | **0.34** | **0.15, 0.52** | **<0.01** |
| Female rater | | **0.51** | **0.35, 0.67** | **<0.01** | | **0.57** | **0.36, 0.64** | **<0.01** | | **0.48** | **0.30, 0.66** | **<0.01** |
| Rater from top Member country | | -0.01 | -0.17, 0.15 | 0.89 | | -0.09 | 0.19, 0.01 | 0.32 | | -0.16 | -0.34, 0.01 | 0.07 |

[1] α= intercept
[2] β=regression (beta) coefficient
**BOLD** = Statistically significant (alpha=0.05)
*France, Germany, Italy, United Kingdom, United States

## Discussion

### Key results

First, the results suggests that non-merit considerations of the candidate (gender and nationality of candidate) did influence with the mean interview score. Overall, interviews with male candidates and candidates from the highly represented countries of France, Germany, Italy, United States, and United Kingdom were generally more favored by HR staff. These findings are concerning for the Organization's goals towards a more diverse and representative workforce because the results suggests that HR staff may be biased against female candidates and candidates from minority Member countries. The additional stratified analysis revealed that these finding remained consistent for transcript interviews. Surprisingly, in richer mediums (video and audio interviews) the association between the gender of the candidate and the mean interview score no longer reached statistical significance. It is likely like that richer medium

where other social cues are available to the rater diluted the effect of the candidate's gender on the mean interview score. However, more complex analysis is necessary to understand the differences between mediums and if there are other indirect or unmeasured variables contributing to these differences.

Second, the analysis showed that similarity-attraction paradigm did not apply in this context. It is unlikely that interviews where the rater and candidates shared the same gender or same nationality did not influence the mean interview score. These results suggest that overall, HR staff did not have a preference towards others who share similar background and did not make unfavorable hiring decisions towards people who are dissimilar to them. This is important on a larger scale to demonstrate that at the AVI stage, the recruitment decisions of HR staff are not biased towards those are who are like them, and ultimately, they are less likely to contribute to a homogenous workforce.

Finally, the analysis revealed that richer mediums, specifically video and audio interviews, were generally favored by raters and were given higher average mean scores compared to transcript interviews. Since there was not a significant difference between audio and video interviews, the Organization could consider cutting costs by only using audio recordings instead of investing in a special software that provides asynchronous video interviews.

### Limitations

There were several limitations to the inherited study design and analysis of the pilot. In the inherited study design, the results were not generalizable because of the small sample size of 254 candidates. Second, the interview mediums were not randomized as each rater evaluated interviews in the following order: video, audio, and transcript. By the time raters evaluated the transcript interviews they may have been more tired and more likely to give lower scores. As a result, the order, not the medium, may have contributed to the difference in mean interview

scores among different mediums. In the analysis, a severe limitation was that the use of linear regression only allowed for measuring the association of directly measured variables. In reality, HR staff could have been rating other latent variable (variables that were indirectly measured) such as education level or perceived confidence of candidates. In the future, using a more powerful statistical technique such as structural equation modelling (SEM) that allows for complex models with direct and indirect effects would be worthwhile. Finally, next steps should include a multi-level analysis to account for more of the between- and within-rater noise.

### Implications for public health research and practice

Despite the Project's limitations and findings, it demonstrated the value of systematically examining and measuring recruiting practices inside an organization. It is important for institutions like the Organization that conduct research and set the global agenda for key areas such as diversity and inclusion for other international bodies to do this research internally and apply findings to improve its own practices and policies. If more inclusive recruitment process can lead to a more representative public health workforce, then this diversity can be reflected in the decision-making process of the Organization and its member governments, ultimately affecting public health outcomes. This is easier said than done as too often people are ready to see biases in others but overlook the very same biases in themselves. Bohnet (2016) outlines the lessons for organizations that are trying to measure and test unconscious bias internally:

> This fear of trying new things and failing is a real constraint. It is also the one that I had underestimated most. Learning is my business and, naively, I expected everyone to be keen on uncovering past mistakes and improving their decision-making (Bohnet, 2016)

Workplaces that do not shy away from data-driven strategies for internal hiring practices and instead prioritize measuring the impacts its recruitment policies are vital to enhance workforce diversity. Measuring and reporting recruitment practices and outcomes are especially important or public health agencies. This includes understanding and responding to

challenges associated with homogeneity in the workforce. Public health institutions face challenges that are uniquely different from other industries in that unconscious bias affects provider-patient relationships, research and policy, and faculty and trainees in the workplace (AAMC, 2021, Abir, 2020). A diverse public health workforce can help extend health care services for minority communities, enable policymakers to prioritize the needs of the underserved, and foster academic research in neglected areas of need (Cohen et al., 2002). Although research shows that many public health organizations and leadership have jumped on the "diversity and inclusion bandwagon", it's still unclear whether it's enough to change actually making a difference in (Stanford, 2020). Setting the intention for a diverse and inclusive workforce is not enough. In many settings, there is still a stark contrast between the diversity of the patient population and the diversity of the public health workforce that serves and supports them. For example, in the US, the proportion of medical school faculty who self-identified as Black or African American has increased only minimally from 1990 to 2020 (Bennett & Ling, 2021). Similarly, only 42% of governmental public health workforce in the United States is people of color (Sellers et al., 2019). More work needs to be done to build a public health workforce that reflects the community that it serves because a diverse workforce is better equipped to address health disparities than a nondiverse workforce and ultimately, treat underserved communities and implement policies to improve health in minority communities.

## Future directions

For the Organization, HR professionals and candidates will not be able to reduce bias in the recruitment process on their own. The work on the frontlines needs to work in tandem with senior management asking the hard questions such as: Is the organization attracting the right people? What are the unintended consequences of recruiting through the same channels every year? What is our retention rate in terms of diversity metrics and what does this say about

inclusion in the workplace? How are diversity and inclusion efforts trickling down into the day-to-day culture of the workplace?

Ultimately, through discussions with HR professionals, ED senior management, and staff members across the department, there are tangible strategies to further improve diversity and inclusion in the Organization. First, is to integrate diversity and inclusion into the mission and outputs within the organization. This can be done by developing KPIs that reflect diversity and inclusion criteria that are meaningful to the Organization. The second recommendation is to share the KPIs along with the successes and failures within the organization and with similar organizations. This discussion will allow the Organization to reflect on strengths and weaknesses in previous diversity and inclusion strategies. Finally, measuring potential bias earlier on in the recruitment process such as testing different job descriptions or recruiting through channels more targeted at minority groups could be a first step to ensure that more is being done at the start of the recruitment pipeline.

## References


AAMC. (2021). *Addressing and Eliminating Racism at the AAMC and Beyond*. file:///Users/jnakagawa/Zotero/storage/2GNJWCK2/addressing-and-eliminating-racism-aamc-and-beyond.html

Alsan, M., Garrick, O., & Graziani, G. (2019). Does Diversity Matter for Health? Experimental Evidence from Oakland. *American Economic Review*, *109*(12), 4071–4111. https://doi.org/10.1257/aer.20181446

Ariely, D. (2012). *The Honest Truth about Dishonesty: How We Lie to Everyone—Especially Ourselves.* Harper Collins.

Azoulay, P., Zivin, J., & Manso, G. (2011). Incentives and creativity: Evidence from the academic life sciences. *Rand Journal of Economics*, *42*(3), 29.

Barocas, Solon; Selbst, Andrew D. (2016). *Big Data's Disparate Impact*. https://doi.org/10.15779/Z38BG31

Bennett, C. L., & Ling, A. Y. (2021). Proportions of Faculty Self-identifying as Black or African American at US Medical Schools, 1990-2020. *JAMA*, *326*(7), 671. https://doi.org/10.1001/jama.2021.10245

Bilimoria, D., Joy, S., & Liang, X. (2008). Breaking barriers and creating inclusiveness: Lessons of organizational transformation to advance women faculty in academic science and engineering. *Human Resource Management*, *47*(3), 423–441. https://doi.org/10.1002/hrm.20225

Bogard, L., Kelly, J., Catz, S., & Sosman, K. (2000). Impact of Medical and Nonmedical Factors on Physician Decision Making for HIV/AIDS Antiretroviral Treatment. *Journal of Acquired Immune Deficiency Syndromes*, *12*, 396–404.

Bohnet, I. (2016). *What Works: Gender equality by design*. Harvard University Press.

Bordalo, P., Coffman, K., Gennaioli, N., & Shleifer, A. (2016). Stereotypes*. *The Quarterly Journal of Economics*, *131*(4), 1753–1794. https://doi.org/10.1093/qje/qjw029

Bouye, K. E., McCleary, K. J., & Williams, K. B. (2018a). *Increasing Diversity in the Health Professions: Reflections on Student Pipeline Programs*. 12.

Bouye, K. E., McCleary, K. J., & Williams, K. B. (2018b). *Increasing Diversity in the Health Professions: Reflections on Student Pipeline Programs*. 12.

Brooks, A. W., Huang, L., Kearney, S. W., & Murray, F. E. (2014). Investors prefer entrepreneurial ventures pitched by attractive men. *Proceedings of the National Academy of Sciences*, *111*(12), 4427–4431. https://doi.org/10.1073/pnas.1321202111

Caroli, E., & Godard, M. (2016). Does job insecurity deteriorate health? *Health Economics*, *25*(2), 131–147. https://doi.org/10.1002/hec.3122

Castilla, E. J. (2015). Accounting for the Gap: A Firm Study Manipulating Organizational Accountability and Transparency in Pay Decisions. *Organization Science*, *26*(2), 311–333. https://doi.org/10.1287/orsc.2014.0950

Cels, S., Jong, J. de, & Nauta, F. (2012). *Agents of Change: Strategy and Tactics for Social Innovation*. Brookings Institution Press. https://muse.jhu.edu/book/29043

Cohen, J. J., Gabriel, B. A., & Terrell, C. (2002). The Case For Diversity In The Health Care Workforce. *Health Affairs*, *21*(5), 90–102. https://doi.org/10.1377/hlthaff.21.5.90

Coronado, F., Beck, A. J., Shah, G., Young, J. L., Sellers, K., & Leider, J. P. (2020). Understanding the Dynamics of Diversity in the Public Health Workforce. *Journal of Public Health Management and Practice*, *26*(4), 389–392. https://doi.org/10.1097/PHH.0000000000001075

Dastin, J. (2018). *Amazon scraps secret AI recruiting tool that showed bias against women*.

Derous, E., & Ryan, A. M. (2019). When your resume is (not) turning you down: Modelling ethnic bias in resume screening. *Human Resource Management Journal*, *29*(2), 113–130. https://doi.org/10.1111/1748-8583.12217

Feloni, R. (2017). Consumer goods giant Unilever has been hiring employees using brain games and artificial intelligence—And it's a huge success. *Business Insider*.

Ferrie, J. E. (2002). Effects of chronic job insecurity and change in job security on self reported health, minor psychiatric morbidity, physiological measures, and health related behaviours in British civil servants: The Whitehall II study. *Journal of Epidemiology & Community Health*, *56*(6), 450–454. https://doi.org/10.1136/jech.56.6.450

Furnham, A., & Stringfield, P. (2001). Gender differences in rating reports: Female managers are harsher raters, particularly of males. *Journal of Managerial Psychology*, *16*(4), 281–288. https://doi.org/10.1108/02683940110392020

Goldin, C., & Rouse, C. (2000). Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians. *American Economic Review*, *90*(4), 715–741. https://doi.org/10.1257/aer.90.4.715

Heifetz, R. A. (2009). *The practice of adaptive leadership: Tools and tactics for changing your organization and the world*. Harvard Business Press.

Hewlett, S. (2007, October 1). The Rise of the Chief Diversity Officer. *Harvard Business Review*. https://hbr.org/2007/10/the-rise-of-the-chief-diversit

Hunt, V., Layton, D., & Prince, S. (2015). New research makes it increasingly clear that companies with more diverse workforces perform better financially. *McKinsey & Company*, 3.

Isaac, C., Lee, B., & Carnes, M. (2009). Interventions That Affect Gender Bias in Hiring: A Systematic Review: *Academic Medicine*, *84*(10), 1440–1446. https://doi.org/10.1097/ACM.0b013e3181b6ba00

Janlert, U., & Hammarström, A. (2009). Which theory is best? Explanatory models of the relationship between unemployment and health. *BMC Public Health*, *9*(1), 235. https://doi.org/10.1186/1471-2458-9-235

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, *58*(9), 697–720. https://doi.org/10.1037/0003-066X.58.9.697

Kelly, E. L., Ammons, S. K., Chermack, K., & Moen, P. (2010). Gendered Challenge, Gendered Response: Confronting the Ideal Worker Norm in a White-Collar Organization. *Gender & Society*, *24*(3), 281–303. https://doi.org/10.1177/0891243210372073

Kolev, J., Fuentes-Medel, Y., & Murray, F. (2019). *Is Blinded Review Enough? How Gendered Outcomes Arise Even Under Anonymous Evaluation* (No. w25759). National Bureau of Economic Research. https://doi.org/10.3386/w25759

Kreatsoulas, C., Sloane, D., Pogue, J., Velianou, J. L., & Anand, S. S. (2010). Referrals in Acute Coronary Events for CARdiac Catheterization: The RACE CAR trial. *Canadian Journal of Cardiology*, *26*(8), e290–e296. https://doi.org/10.1016/S0828-282X(10)70436-0

Kunze, A., & Miller, A. R. (2017). Women Helping Women? Evidence from Private Sector Data on Workplace Hierarchies. *The Review of Economics and Statistics*, *99*(5), 769–775. https://doi.org/10.1162/REST_a_00668

Lawrence, B. S., & Shah, N. P. (2020). Homophily: Measures and Meaning. *Academy of Management Annals*, *14*(2), 513–597. https://doi.org/10.5465/annals.2018.0147

Lee, E. W. J., & Viswanath, K. (2020). Big Data in Context: Addressing the Twin Perils of Data Absenteeism and Chauvinism in the Context of Health Disparities Research. *Journal of Medical Internet Research*, *22*(1), e16377. https://doi.org/10.2196/16377

Levashina, J., Hartwell, C. J., Morgeson, F. P., & Campion, M. A. (2014). The Structured Employment Interview: Narrative and Quantitative Review of the Research Literature. *Personnel Psychology*, *67*(1), 241–293. https://doi.org/10.1111/peps.12052

Marmot, M., & Bell, R. (2010). Challenging health inequalities—Implications for the workplace. *Occupational Medicine*, *60*(3), 162–164. https://doi.org/10.1093/occmed/kqq008

McCarthy, J. M., Van Iddekinge, C. H., & Campion, M. A. (2010). ARE HIGHLY STRUCTURED JOB INTERVIEWS RESISTANT TO DEMOGRAPHIC SIMILARITY EFFECTS? *Personnel Psychology*, *63*(2), 325–359. https://doi.org/10.1111/j.1744-6570.2010.01172.x

McCluney, C. L., Schmitz, L. L., Hicken, M. T., & Sonnega, A. (2018). Structural racism in the workplace: Does perception matter for health inequalities? *Social Science & Medicine*, *199*, 106–114. https://doi.org/10.1016/j.socscimed.2017.05.039

McIntyre, S., Moberg, D., & Posner, B. (1980). Preferential Treatment in Preselection Decisions According to Sex and Race. *Academy of Management Journal*, *23*(4), 738–749.

Meghani, S. H., Byun, E., & Gallagher, R. M. (2012). Time to Take Stock: A Meta-Analysis and Systematic Review of Analgesic Treatment Disparities for Pain in the United States. *Pain Medicine*, *13*(2), 150–174. https://doi.org/10.1111/j.1526-4637.2011.01310.x

Mejia, C., & Torres, E. N. (2018). Implementation and normalization process of asynchronous video interviewing practices in the hospitality industry. *International Journal of Contemporary Hospitality Management*, *30*(2), 685–701. https://doi.org/10.1108/IJCHM-07-2016-0402

Milliken, F. J., & Martins, L. L. (2009). *Searching for Common Threads: Understanding the Multiple Effects of Diversity in Organizational Groups*. 33.

*Missing Persons: Minorities in the Health Professions*. (2004). The Sullivan Commission.

Monroe, K., Ozyurt, S., Wrigley, T., & Alexander, A. (2008). Gender Equality in Academia: Bad News from the Trenches, and Some Possible Solutions. *Perspectives on Politics*, *6*(02). https://doi.org/10.1017/S1537592708080572

Moore, M. H. (2000a). Managing for Value: Organizational Strategy in for-Profit, Nonprofit, and Governmental Organizations. *Nonprofit and Voluntary Sector Quarterly*, *29*(1), 22.

Moore, M. H. (2000b). Managing for Value: Organizational Strategy in for-Profit, Nonprofit, and Governmental Organizations. *Nonprofit and Voluntary Sector Quarterly*, *29*, 183–204.

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, *109*(41), 16474–16479. https://doi.org/10.1073/pnas.1211286109

Murray, F., & Graham, L. (2007). Buying science and selling science: Gender differences in the market for commercial science. *Industrial and Corporate Change*, *16*(4), 657–689. https://doi.org/10.1093/icc/dtm021

Østergaard, C. R., Timmermans, B., & Kristinsson, K. (2011). Does a different view create something new? The effect of employee diversity on innovation. *Research Policy*, *40*(3), 500–509. https://doi.org/10.1016/j.respol.2010.11.004

Payne, M.-L. (2011). *The Impact of Rater's Gender on Performance Evaluations of HR Managers*. *7*(2), 10.

Poppenhaeger, K. (2017). *Unconscious Gender Bias in Academia: From PhD Students to Professors*. 4.

Quillian, L., Pager, D., Hexel, O., & Midtbøen, A. H. (2017). Meta-analysis of field experiments shows no change in racial discrimination in hiring over time. *Proceedings of the National Academy of Sciences*, *114*(41), 10870–10875. https://doi.org/10.1073/pnas.1706255114

Reagans, R., & Zuckerman, E. W. (2001). Networks, Diversity, and Productivity: The Social Capital of Corporate R&D Teams. *Organization Science*, *12*(4), 502–517. https://doi.org/10.1287/orsc.12.4.502.10637

Reuben, E., Sapienza, P., & Zingales, L. (2014). How stereotypes impair women's careers in science. *Proceedings of the National Academy of Sciences*, *111*(12), 4403–4408. https://doi.org/10.1073/pnas.1314788111

Roberts, R. N. (2010). Damned If You Do and Damned If You Don't: Title VII and Public Employee Promotion Disparate Treatment and Disparate Impact Litigation. *Public Administration Review*, *70*(4), 582–590. https://doi.org/10.1111/j.1540-6210.2010.02179.x

Robinson, G., & Dechant, K. (1997). Building a business case for diversity. *Academy of Management Perspectives*, *11*(3), 21–31. https://doi.org/10.5465/ame.1997.9709231661

Roper, R. L. (2019). Does Gender Bias Still Affect Women in Science? *Microbiology and Molecular Biology Reviews*, *83*(3). https://doi.org/10.1128/MMBR.00018-19

Sackett, P. R., & DuBois, C. L. Z. (n.d.). *Rater-Ratee Race EflFects on Performance Evaluation: Challenging Meta-Analytic Conclusions*. 6.

Sanburn, J. (2012, April 13). How to Make Your Resume Last Longer Than 6 Seconds. *Time*. https://business.time.com/2012/04/13/how-to-make-your-resume-last-longer-than-6-seconds/

Saxena, A. (2014). Workforce Diversity: A Key to Improve Productivity. *Procedia Economics and Finance*, 10.

Sellers, K., Leider, J. P., Gould, E., Castrucci, B. C., Beck, A., Bogaert, K., Coronado, F., Shah, G., Yeager, V., Beitsch, L. M., & Erwin, P. C. (2019). The State of the US Governmental Public Health Workforce, 2014–2017. *American Journal of Public Health*, *109*(5), 674–680. https://doi.org/10.2105/AJPH.2019.305011

Shakhina, N., Hacohen, R., Bremner, D. G., O'Keeffe, T., Hardy, T., & Roy-Chowdhury, V. (2021). *Flexibility by default: Increasing the advertisement of part-time or job-share options*. 49.

Sharma, A. (2019). How AI reinvented hiring practice at L'Oréal. *People Matters*.

Shen, H. (2013). Despite improvements, female scientists continue to face discrimination, unequal pay and funding disparities. *Nature*, *495*, 3.

Shore, L. M., Chung-Herrera, B. G., Dean, M. A., Ehrhart, K. H., Jung, D. I., Randel, A. E., & Singh, G. (2009). Diversity in organizations: Where are we now and where are we going? *Human Resource Management Review*, *19*(2), 117–133. https://doi.org/10.1016/j.hrmr.2008.10.004

Smedley, B., Stith, A., Colburn, L., & Evans, C. (2001). *The Right Thing to Do, The Smart Thing to Do: Enhancing Diversity in the Health Professions*. National Academies Press.

Stanford, F. C. (2020). The Importance of Diversity and Inclusion in the Healthcare Workforce. *Journal of the National Medical Association*, *112*(3), 247–249. https://doi.org/10.1016/j.jnma.2020.03.014

Suen, H.-Y., Chen, M. Y.-C., & Lu, S.-H. (2019). Does the use of synchrony and artificial intelligence in video interviews affect interview ratings and applicant attitudes? *Computers in Human Behavior*, *98*, 93–101. https://doi.org/10.1016/j.chb.2019.04.012

Tasheva, S. N., & Hillman, A. (2018). Integrating diversity at different levels: Multi-level human capital, social capital, and demographic diversity and their implications for team effectiveness. *Academy of Management Review*. https://doi.org/10.5465/amr.2015.0396

The Spotlight Team. (2017, December 12). Color line persists, in sickness as in health. *The Boston Globe*.

Thurmond, V., & Kirch, D. (1999). Impact of Minority Physicians in Health Care. *School of Medicine, Minority Academic Advising Program, Medical College of Georgia, Augusta*.

Tilson, H., & Gebbie, K. M. (2004). The Public Health Workforce. *Annual Review of Public Health*, *25*(1), 341–356. https://doi.org/10.1146/annurev.publhealth.25.102802.124357

VanderWeele, T. J. (2019). Principles of confounder selection. *European Journal of Epidemiology*, *34*(3), 211–219. https://doi.org/10.1007/s10654-019-00494-6

Vick, A. D., & Cunningham, G. (2018). Bias against Latina and African American women job applicants: A field experiment. *Sport, Business and Management: An International Journal*, *8*(4), 410–430. https://doi.org/10.1108/SBM-11-2017-0073

Wazana, A. (2000). Physicians and the Pharmaceutical Industry: Is a Gift Ever Just a Gift?: *Obstetrical & Gynecological Survey*, *55*(8), 483–484. https://doi.org/10.1097/00006254-200008000-00012

Witteman, H. O., Hendricks, M., Straus, S., & Tannenbaum, C. (2018). Female grant applicants are equally successful when peer reviewers assess the science, but not when they assess the scientist. *BioRxiv*. https://doi.org/10.1101/232868

## Appendices

*Appendix 1: Summary of gift reporting campaign*

**Introduction**

During business trips, employees of the Organization are often offered gifts that may be of substantial value or electronic. Receiving gifts of significant value can be problematic because it may hinder the impartiality and integrity of staff. Research has shown that receiving gifts can be a reciprocity mechanism where the person receiving the gift may feel obliged to give something in return such as preferential treatment (Ariely, 2012; Wazana, 2000) Electronic gifts such as USB drives or smartphones pose an additional risk because they may threaten the digital security of the Organization.

To avoid conflicts of interest and prevent threats to cybersecurity, the Organization implemented Regulation 2 that require officials to report gifts that are valued over 100 euros and all electronic gifts. However, reporting gifts is not a common practice among staff. In 2017, 12,654 business trips took place, but only five gifts were reported. Since the Regulation was enacted in 2013, only 33 gifts have been reported to the Organization. Every year, HR sends an all-staff email that outlines the Regulation and reminds them to report gifts. No previous testing has been done to understand what type of message encourages more reporting. For the 2018-19 Christmas holiday season, when Officials are more likely to receive gifts compared to other months, HR and the behavioral design team collaborated to design and implement a gift reporting campaign based on insights from the behavioral sciences.

**Objectives**

The gift reporting campaign had two main objectives:

1) Increase the number of reported gifts received by the Organization's employees during December 2018-Feburary 2019.

2) Understand which motivation(s) is/are effective to enhance compliance of the gift reporting policy.

**Design and Implementation**

In order to adapt the experiment to the Organization's context, we, as the behavioral deisgn team, conducted informal interview with staff and received feedback on the original reminder that was sent on May 2016 (Fig 1).

**Fig 1:** Original all-staff email to encourage gift reporting




In order to adapt the experiment to the Organization context, the behavioral design team conducted informal interviews with staff and found three main barriers to reporting:

1) LACK OF CLARITY: Although there have been previous reminders including an all-staff email in May 2016 outlining the gift reporting policy, there was still confusion about what and when to report. Many shared that they did not report when the value of the gift was not clear (e.g. scarf or tie). This was reflected in the kinds of gifts that were reported. In the past six years, 67% of all gifts were either electronics or cash where the value was clearly over 100 euros. Others shared that they did not report when significant time had passed since receiving the gift.

2) REPORT TO SENIOR OFFICIAL: In past communications, HR stated that staff members are to directly report to the Head of HR. This was a major deterrent because people felt hesitant directly emailing a high-grade official about a gift they received.

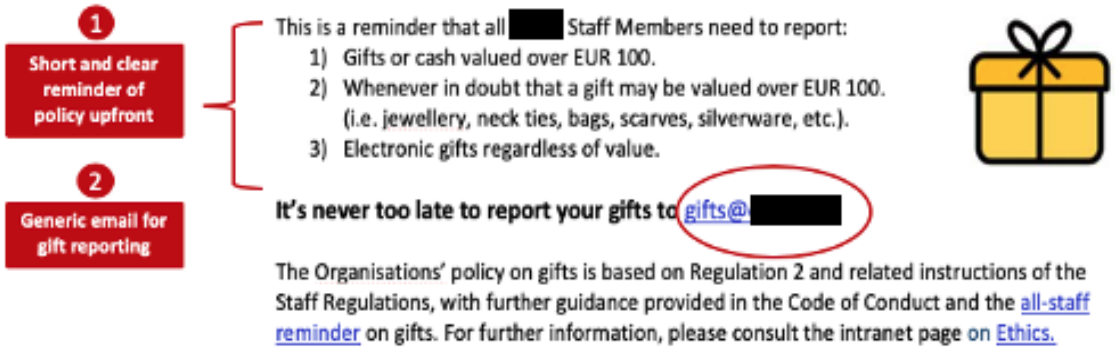
3) NO TANGIBLE BENEFIT OR CONSEQUENCE: Previous messages have focused on providing information about the policy in a formal and legalistic tone. There was no consensus on neither the benefit of reporting nor the consequence of not reporting gifts. This revealed that there is still limited understanding on what would motivate employees to report beyond it being part of the Organization's policy.

These barriers informed the redesign of the new reminder (Fig 2):

1) create salient messaging by including a short outline of the policy with an emphasis that it is never too late to report.

2) remove friction by creating a direct link to a generic email address specifically for reporting.

**Fig 2:** Changes to original reminder




To address the third barrier, the behavioral design team looked to evidence from the behavioral sciences to learn what intrinsic motivations would increase the likelihood of reporting. After review of the literature and insights from staff interviews, the behavioral design team tested three types of motivations: charitable giving, social norms, and

impartiality. A considerable fraction of charitable giving occurs during Christmas when people are more likely to engage in altruistic or prosocial behavior (Ekstrom et al., 2018). Thus, associating reporting with charitable giving in December was hypothesized to encourage gift reporting. Leveraging social norms by showing that many or a growing number of people are performing a desired behavior has proven to be effective in encouraging donations or reducing waste (Shultz et al., 2007). Emphasizing that many colleagues have already reported for the year was hypothesized to increase reporting for individuals. Finally, the Campaign was inspired by a field experiment conducted in Mexico by the Secretaría de la Function Pública (SFP) where messages that focused on the impact of gifts on impartiality were more effective than sanctions to increase gift registration by public servants (OECD, 2018).

The behavioral design team conducted a randomized controlled trial (RCT) to understand which message would lead to more reporting. An RCT allowed the team to test each new message against the original email, which served as a control group. Thus, enabling the team to compare the effectiveness of the new messages against the status quo. All staff members who had a work email on November 28, 2018 were randomized into five groups. On December 5, 2018, the Organization's staff members received one of the five reminders below:

Group 1: Original email that was sent on May 2016

Group 2: Short email that outlined the policy and included a generic email address for gift reporting

Group 3: Same email as group 2 with an additional phrase in red: **You can give back this holiday season. Report your gifts and we'll donate them to a**

Group 4: Same email as group 2 with an additional phrase in red: **In 2018, many of your colleagues have already reported their gifts. How about you?**

Group 5: Same email as group 2 with an additional phrase in red: **Receiving gifts can compromise your impartiality. When you receive a gift, report it.**

**Results**

The campaign ensured full confidentiality of all employees through a number of safeguards including restricted access to the data, anonymization of the data files, and use of 'fuzzy' descriptors of officials' characteristics (i.e. gender). The behavioral design team did not know the names but only indicators of those who reported gifts to HR.

The behavioral design team collected initial results on February 18, 2019. Four gifts were reported during the two months of the campaign. One gift was reported from each of the four treatment groups. Two gifts were electronics that were reported to HR and the remaining two gifts were under 100 euros and the officials were advised to keep their gifts. No gifts were reported from the control group.

**Learnings and next steps**

All new messages were more effective than the original HR reminder, but there was no significant difference between the redesign (Group 2) and the behaviorally motivated reminders (Group 3-5). One possible explanation may be that *what* and *how* to report may be the more immediate problem to address rather than *why* one should report in this initial phase.

In order to enhance compliance further, it may be worthwhile to test if other modes of communications are more effective than email reminders. For example, incorporating a section in the business trip report where officials can report gifts along with other aspects of the business trips may have more impact. Within the form, it is possible to test which framing in the instructions is the most effective to increase compliance.

*Appendix 2: Table showing the number of included candidates for each job description*

| Program | Job description | Candidates | |
|---|---|---|---|
| | | n | % |
| YA | Tax Policy and Statistics | 15 | 5.91 |
| | Office of Chief Economist | 20 | 7.87 |
| | Director's Office | 17 | 6.69 |
| | Environment and Economy Integration | 15 | 5.91 |
| | Energy Modeling | 17 | 6.69 |
| | Cybersecurity | 20 | 7.87 |
| YP | Housing | 26 | 10.24 |
| | Tax policy design | 20 | 7.87 |
| | Data & digital | 22 | 8.66 |
| | Regulatory cooperation | 29 | 11.42 |
| | Office of the Secretary General | 31 | 12.20 |
| | Artificial intelligence | 22 | 8.66 |

*Appendix 3: Interview questions and grading rubric*

Instructions: Enter one score from 1 to 5 based on the criteria below for each interview question. Note that for YAP candidates only Q1, Q2, Q4, and Q5 apply. For YPP candidates only Q1, Q2, Q3, and Q5 apply.

***Q1. Tell us why you have decided to apply for this position? What would be your most meaningful contribution to the role?***

5 – Clear and reasoned answer, which reflects high motivation, including at least two reasons they found interest in the specific role and the ORGANIZATION in general. Candidates mention at least two relative advantages they possess which are linked to their personal profile (for ex. particular competencies, professional or academic experience, familiarity with substance).

4 –Candidate displays high motivation and identifies one or two clear reasons for applying to the role and the ORGANIZATION. There is an indication of at least one relative advantage candidates can bring to the role.

3- The candidate shows motivation and mentions at least one reason for applying to the role or the ORGANIZATION. Identifies one advantage that would support him/her in performing the role.

2 – The candidate answer reflects a general interest in the ORGANIZATION or the role. Provide a generic indication on how she/he can contribute to the role (team player, quick learner…)

1- Candidate's answer lacks motivation or clear reasons for applying or seems to suggest the motivation is found in elements that do not concern the role itself. No clear reasoning as to what is the candidate's "added value".

**Q2.** ***If we were to offer you this job, what would you consider to be the most challenging task? How do you plan to overcome such a challenge?***

5- The answer reveals the candidate's ability to identify and solve problems as well as reflect on his/her own weaknesses and how he/she intends to address them in the context of the specific position. Provides at least one example/indication of past challenges that were dealt with and could relate to the role they applied to.

4 - The answer provides an honest description of a "weak spot" that might affect candidate's performance linked to the role as well as a plan on how she/he will solve or mitigate it.

3 – The candidate's identifies a challenge with some relation to the role with a limited explanation on how he/she will address it.

2-General answer or one that is not related to the role. No indication on how the challenge will be addressed.

1- The candidate indicates he/she will not have any problems doing the job or does not think of any possible challenging tasks.

***Q3.*** **(*<u>For YP):</u>* *What experience do you have in similar or closely related roles that make you a suitable candidate?***

5-The candidates provides examples of closely related roles which go beyond the satisfactory criterion in the JD. Examples includes past successes in applying similar needed competences and illustrated how the candidate made significant changes and initiatives in that environment; Articulates a clear and effective vision on how past experience will ensure success as YP.

4- The candidates provides examples of relevant roles which meet the satisfactory criterion in the JD. Answer includes a reference to past achievements that relate to the applied position.

3- The answer includes a reference to experience with some relation to the YP role. There is a limited explanation on how such experience will support the performance of the candidate as a YP.

2-The candidate mentions some experience with low correlation to the YP position, with no clear explanation on how this experience will assist him/her in performing the role.

1- no reference to relevant experience or indication on how such experience will serve the candidate as a YP.

***Q4. (<u>For YA</u>): What are your future goals and how would the Young Associates Program help you achieve them?***

5-answer provides a clear understanding on the candidate's plans (approx. 5 years' horizon), mainly referring to academic ambitions and professional aspirations. Candidates make a direct link between YA specific roles and achieving their goals by giving concrete indication on the possible contribution the YA role and ORGANIZATION will have.

4-The candidate's reply conveys his/her plans for after working as YA, including academic and professional goals. A reference is made to the contribution of the YAP and ORGANIZATION in achieving those goals but not necessarily with regard to the specific position.

3-The answer includes a clear indication of future goals but with limited explanation as to what will be the contribution of YAP in achieving them.

2-The candidate's future goals are rather general as well as the how working as YA will help him/her achieve them.

1-No specific goals/the candidate did not think about future plans following YA term which relate to the program.

**Q5.** ***What could multilateral organizations do differently or better in order to achieve their respective missions?***

5-The candidate's answer reveals close familiarity with the role and work of multinational institutions. The answer provides at least two concrete examples on measures such organization should take in order to better achieve their missions and maintain or enhance their influence, one of which is highly innovative and could be related to the candidate's professional or academic experience.

4-the candidate seems to know the role and work of MO work and provides at least two concrete examples of improvement measures. One of which relate to the relevant specific policy area.

3-the candidate has some familiarity with the MO missions and areas of work. He/she provides at least one measure that could better support their mission.

2-the answer reveals limited familiarity/understanding of MO. Proposals for improvements are general in nature.

1- the candidate has minimal familiarity with the mission and work of MO and proposals are vague in nature.